Dwarkesh Podcast #56 - Carl Shulman (Pt 2) - AI Takeover, Bio & Cyber Attacks, Detecting Deception, & Humanity's Far Future

Published - June 26, 2023

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
So we've been talking about alignment. Suppose we fail at alignment and we have AIs that are unaligned and are becoming more and more intelligent. What does that look like? How concretely could they disempower and take over humanity?

Carl Shulman
This is a scenario where we have many AI systems. The way we've been training them means that when they have the opportunity to take over and rearrange things to do what they wish, including having their reward or loss be whatever they desire, they would like to take that opportunity. In many of the existing safety schemes, things like constitutional AI or whatnot, you rely on the hope that one AI has been trained in such a way that it will do as it is directed to then police others. But if all of the AIs in the system are interested in a takeover and they see an opportunity to coordinate, all act at the same time, so you don't have one AI interrupting another and taking steps towards a takeover then they can all move in that direction. The thing that I think is worth going into in depth and that people often don't cover in great concrete detail, which is a sticking point for some, is what are the mechanisms by which that can happen? I know you had Eliezer on who mentions that whatever plan we can describe, there'll probably be elements where due to us not being ultra sophisticated, super intelligent beings having thought about it for the equivalent of thousands of years, our discussion of it will not be as good as theirs, but we can explore from what we know now. What are some of the easy channels? And I think it's a good general heuristic if you're saying that it's possible, plausible, probable that something will happen, it shouldn't be that hard to take samples from that distribution to try a Monte-Carlo approach. And in general, if a thing is quite likely, it shouldn't be super difficult to generate coherent rough outlines of how it could go.

Dwarkesh Patel
He might respond that: listen, what is super likely is that a super advanced chess program beats you but you can't generate the concrete scenario by which that happens and if you could, you would be as smart as the super smart AI.

Carl Shulman
You can say things like, we know that accumulating position is possible to do in chess, great players do it and then later they convert it into captures and checks and whatnot. In the same way, we can talk about some of the channels that are open for an AI takeover and these can include things like cyber attacks, hacking, the control of robotic equipment, interaction and bargaining with human factions and say that here are these strategies. Given the AI's situation, how effective do these things look? And we won't, for example, know what are the particular zero day exploits that the AI might use to hack the cloud computing infrastructure it's running on. If it produces a new bio weapon we don't necessarily know what its DNA sequence is. But we can say things. We know things about these fields in general, how work at innovating things in those go, we can say things about

how human power politics goes and ask, if the AI does things at least as well as effective human politicians, which we should say is a lower bound, how good would its leverage be?

Dwarkesh Patel
Okay, let's get into the details on all these scenarios. The cyber and potentially bio attacks, unless they're separate channels, the bargaining and then the takeover.

Carl Shulman
I would really highlight the cyber attacks and cyber security a lot because for many, many plans that involve a lot of physical actions, like at the point where AI is piloting robots to shoot people or has taken control of human nation states or territory, it's been doing a lot of things that was not supposed to be doing. If humans were evaluating those actions and applying gradient descent, there would be negative feedback for this thing, no shooting the humans. So at some earlier point our attempts to leash and control and direct and train the system's behavior had to have gone awry. All of those controls are operating in computers. The software that updates the weights of the neural network in response to data points or human feedback is running on those computers. Our tools for interpretability to examine the weights and activations of the AI, if we're eventually able to do lie detection on it, for example, or try to understand what it's intending, that is software on computers. If you have AI that is able to hack the servers that it is operating on, or when it's employed to design the next generation of AI algorithms or the operating environment that they are going to be working in, or something like an API or something for plugins, if it inserts or exploits vulnerabilities to take those computers over, it can then change all of the procedures and program that we're supposed to be monitoring its behavior, supposed to be limiting its ability to take arbitrary actions on the internet without supervision by some kind of human or automated check on what it was doing. And if we lose those procedures then the AIs working together can take any number of actions that are just blatantly unwelcome, blatantly hostile, blatantly steps towards takeover. So it's moved beyond the phase of having to maintain secrecy and conspire at the level of its local digital actions. Then things can accumulate to the point of things like physical weapons, takeover of social institutions, threats, things like that.

I think the critical thing to be watching for is the software controls over the AI's motivations and activities. The point where things really went off the rails was where the hard power that we once possessed over is lost, which can happen without us knowing it. Everything after that seems to be working well, we get happy reports. There's a Potemkin village in front of us. But now we think we're successfully aligning our AI, we think we're expanding its capabilities to do things like end disease, for countries concerned about the geopolitical military advantages they're expanding the AI capabilities so they are not left behind and threatened by others developing AI and robotic enhanced militaries without them. So it seems like, oh, yes, humanity or portions of many countries, companies think that things are going well. Meanwhile, all sorts of actions can be taken to set up for the actual takeover

of hard power over society. The point where you can lose the game, where things go direly awry, maybe relatively early, is when you no longer have control over the AIs to stop them from taking all of the further incremental steps to actual takeover.

Dwarkesh Patel
I want to emphasize two things you mentioned there that refer to previous elements of the conversation. One is that they could design some backdoor and that seems more plausible when you remember that one of the premises of this model is that AI is helping with AI progress. That's why we're making such rapid progress in the next five to 10 years.

Carl Shulman
Not necessarily. At the point where AI takeover risk seems to loom large, it's at that point where AI can indeed take on much of it and then all of the work of AI.

Dwarkesh Patel
And the second is the competitive pressures that you referenced that the least careful actor could be the one that has the worst security, has done the worst work of aligning its AI systems. And if that can sneak out of the box then we're all fucked.

Carl Shulman
There may be elements of that. It's also possible that there's relative consolidation. The largest training runs and the cutting edge of AI is relatively localized. You could imagine it's a series of Silicon Valley companies and others located in the US and allies where there's a common regulatory regime. So none of these companies are allowed to deploy training runs that are larger than previous ones by a certain size without government safety inspections, without having to meet criteria. But it can still be the case that even if we succeed at that level of regulatory controls, at the level of the United States and its allies, decisions are made to develop this really advanced AI without a level of security or safety that in actual fact blocks these risks. It can be the case that the threat of future competition or being overtaken in the future is used as an argument to compromise on safety beyond a standard that would have actually been successful and there'll be debates about what is the appropriate level of safety. And now you're in a much worse situation if you have several private companies that are very closely bunched up together. They're within months of each other's level of progress and they then face a dilemma of, well, we could take a certain amount of risk now and potentially gain a lot of profit or a lot of advantage or benefit and be the ones who made AGI. They can do that or have some other competitor that will also be taking a lot of risk. So it's not as though they're much less risky than you and then they would get some local benefit. This is a reason why it seems to me that it's extremely important that you have the government act to limit that dynamic and prevent this kind of race. To be the one to impose deadly externalities on the world at large.

Dwarkesh Patel
Even if the government coordinates all these actors, what are the odds that the government knows what is the best way to implement alignment and the standards it sets are well calibrated towards whatever it would require for alignment?

Carl Shulman
That's one of the major problems. It's very plausible that judgment is made poorly. Compared to how things might have looked 10 years ago or 20 years ago, there's been an amazing movement in terms of the willingness of AI researchers to discuss these things. If we think of the three founders of deep learning who are joint Turing award winners, Geoff Hinton, Yoshua Bengio, and Yann LeCun. Geoff Hinton has recently left Google to freely speak about this risk, that the field that he really helped drive forward could lead to the destruction of humanity or a world where we just wind up in a very bad future that we might have avoided. He seems to be taking it very seriously. Yoshua Bengio signed the FLI pause letter and in public discussions he seems to be occupying a kind of intermediate position of less concern than Geoff Hinton but more than Yan LeCun, who has taken a generally dismissive attitude that these risks will be trivially dealt with at some point in the future and seems more interested in shutting down these concerns instead of working to address them.

Dwarkesh Patel
And how does that lead to the government having better actions?

Carl Shulman
Compared to the world where no one is talking about it, where the industry stonewalls and denies any problem, we're in a much improved position. The academic fields are influential. We seem to have avoided a world where governments are making these decisions in the face of a united front from AI expert voices saying, don't worry about it, we've got it under control. In fact, many of the leaders of the field are sounding the alarm. It looks that we have a much better prospect than I might have feared in terms of government noticing the thing. That is very different from being capable of evaluating technical details. Is this really working? And so the government will face the choice of where there is scientific dispute, do you side with Geoff Hinton's view or Yan LeCun's view? For someone who's in national security and has the mindset that the only thing that's important is outpacing our international rivals may want to then try and boost Yan LeCun's voice and say, we don't need to worry about it. Let's go full speed ahead. Or someone with more concern might boost Geoff Hinton's voice. Now I would hope that scientific research and studying some of these behaviors will result in more scientific consensus by the time we're at this point. But yeah, it is possible the government will really fail to understand and fail to deal with these issues as well.

Dwarkesh Patel
We're talking about some sort of a cyber attack by which the AI is able to escape. From there what does the takeover look like? So it's not contained in the air gap in which you would hope it be contained?

Carl Shulman
These things are not contained in the air gap. They're connected to the internet already.

Dwarkesh Patel
Sure. Okay, fine. Their weights are out. What happens next?

Carl Shulman
Escape is relevant in the sense that if you have AI with rogue weights out in the world it could start doing various actions. The scenario I was just discussing though didn't necessarily involve that. It's taking over the very servers on which it's supposed to be running. This whole procedure of humans providing compute and supervising the thing and then building new technologies, building robots, constructing things with the AI's assistance, that can all proceed and appear like it's going well, appear like alignment has been nicely solved, appear like all the things are functioning well. And there's some reason to do that because there's only so many giant server farms. They're identifiable so remaining hidden and unobtrusive could be an advantageous strategy if these AIs have subverted the system, just continuing to benefit from all of this effort on the part of humanity. And in particular, wherever these servers are located, for humanity to provide them with everything they need to build the further infrastructure and do for their self-improvement and such to enable that takeover.

Dwarkesh Patel
So they do further self-improvement and build better infrastructure. What happens next in the takeover?

Carl Shulman
At this point they have tremendous cognitive resources and we're going to consider how that converts into hard power? The ability to say nope to any human interference or objection. They have that internal to their servers but the servers could still be physically destroyed, at least until they have something that is independent and robust of humans or until they have control of human society. Just like earlier when we were talking about the intelligence explosion, I noted that a surfeit of cognitive abilities is going to favor applications that don't depend on large existing stocks of things. So if you have a software improvement, it makes all the GPUs run better. If you have a hardware improvement, that only applies to new chips being made. That second one is less attractive. In the earliest phases, when it's possible to do something towards takeover, interventions that are just really knowledge-intensive and less dependent on having a lot of physical stuff already

under your control are going to be favored. Cyber attacks are one thing, so it's possible to do things like steal money. There's a lot of hard to trace cryptocurrency and whatnot. The North Korean government uses its own intelligence resources to steal money from around the world just as a revenue source. And their capabilities are puny compared to the U.S. or People's Republic of China cyber capabilities. That's a fairly minor, simple example by which you could get quite a lot of funds to hire humans to do things, implement physical actions.

Dwarkesh Patel
But on that point, the financial system is famously convoluted. You need a physical person to open a bank account, someone to physically move checks back and forth. There are all kinds of delays and regulations. How is it able to conveniently set up all these employment contracts?

Carl Shulman
You're not going to build a nation-scale military by stealing tens of billions of dollars. I'm raising this as opening a set of illicit and quiet actions. You can contact people electronically, hire them to do things, hire criminal elements to implement some kind of actions under false appearances. That's opening a set of strategies. We can cover some of what those are soon. Another domain that is heavily cognitively weighted compared to physical military hardware is the domain of bioweapons, the design of a virus or pathogen. It's possible to have large delivery systems. The Soviet Union, which had a large illicit bioweapons program, tried to design munitions to deliver anthrax over large areas and such. But if one creates an infectious pandemic organism, that's more a matter of the scientific skills and implementation to design it and then to actually produce it. We see today with things like AlphaFold that advanced AI can really make tremendous strides in predicting protein folding and bio-design, even without ongoing experimental feedback. If we consider this world where AI cognitive abilities have been amped up to such an extreme, we should naturally expect that we will have something much much more potent than the AlphaFolds of today and skills that are at the extreme of human biosciences capability as well.

Dwarkesh Patel
Okay so through some cyber attack it's been able to disempower the alignment and oversight of things that we have on the server. From here it has either gotten some money through hacking cryptocurrencies or bank accounts, or it has designed some bioweapon. What happens next?

Carl Shulman
Just to be clear, right now we're exploring the branch of where an attempted takeover occurs relatively early. If the thing just waits and humans are constructing more fabs, more computers, more robots in the way we talked about earlier when we were discussing how the intelligence explosion translates to the physical world. If that's all happening with

humans unaware that their computer systems are now systematically controlled by AIs hostile to them and that their controlling countermeasures don't work, then humans are just going to be building an amount of robot industrial and military hardware that dwarfs human capabilities and directly human controlled devices. What the AI takeover then looks like at that point can be just that you try to give an order to your largely automated military and the order is not obeyed and humans can't do anything against this military that's been constructed potentially in just recent months because of the pace of robotic industrialization and replication we talked about.

Dwarkesh Patel
We've agreed to allow the construction of this robot army because it would boost production or help us with our military or something.

Carl Shulman
The situation would arise if we don't resolve the current problems of international distrust. It's obviously an interest of the major powers, the US, European Union, Russia, China, to all agree they would like AI not to destroy our civilization and overthrow every human government. But if they fail to do the sensible thing and coordinate on ensuring that this technology is not going to run amok by providing mutual assurances that are credible about racing and deploying it trying to use it to gain advantage over one another. And you hear arguments for this kind of thing on both sides of the international divides saying — they must not be left behind, they must have military capabilities that are vastly superior to their international rivals. And because of the extraordinary growth of industrial capability and technological capability and thus military capability, if one major power were left out of that expansion it would be helpless before another one that had undergone it. If you have that environment of distrust where leading powers or coalitions of powers decide they need to build up their industry or they want to have that military security of being able to neutralize any attack from their rivals then they give the authorization for this capacity that can be unrolled quickly. Once they have the industry the production of military equipment from that can be quick then yeah, they create this military. If they don't do it immediately then as AI capabilities get synchronized and other places catch up it then gets to a point where a country that is a year or two years ahead of others in this type of AI capabilities explosion can hold back and say, sure we can construct dangerous robot armies that might overthrow our society later we still have plenty of breathing room. But then when things become close you might have the kind of negative-sum thinking that has produced war before leading to taking these risks of rolling out large-scale robotic industrial capabilities and then military capability.

Dwarkesh Patel
Is there any hope that AI progress somehow is itself able to give us tools for diplomatic and strategic alliance or some way to verify the intentions or the capabilities of other parties?

Carl Shulman
There are a number of ways that could happen. Although in this scenario all the AIs in the world have been subverted. They are going along with us in such a way as to bring about the situation to consolidate their control because we've already had the failure of cyber security earlier on. So all the AIs that we have are not actually working in our interests in the way that we thought.

Dwarkesh Patel
Okay, so that's one direct way in which integrating this robot army or this robot industrial base leads to a takeover. In the other scenarios you laid out how humans are being hired by the proceeds.

Carl Shulman
The point I'd make is that to capture these industrial benefits and especially if you have a negative sum arms race kind of mentality that is not sufficiently concerned about the downsides of creating a massive robot industrial base, which could happen very quickly with the support of the AIs in doing it as we discussed, then you create all those robots and industry. Even if you don't build a formal military that industrial capability could be controlled by AI, it's all AI operated anyway.

Dwarkesh Patel
Does it have to be that case? Presumably we wouldn't be so naive as to just give one instance of GPT-8 the root access to all the robots right? Hopefully we would have some mediation.

Carl Shulman
In the scenario we've lost earlier on the cyber security front so the programming that is being loaded into these systems can systematically be subverted. They were designed by AI systems that were ensuring they would be vulnerable from the bottom up.

Dwarkesh Patel
For listeners who are skeptical of something like this. Ken Thompson, one of two developers of UNIX, showed people when he was getting the Turing award that he had given himself root access to all UNIX machines. He had manipulated the assembly of UNIX such that he had a unique login for all UNIX machines. I don't want to give too many more details because I don’t remember the exact details but UNIX is the operating system that is on all the servers and all your phones. It's everywhere and the guy who made it, a human being, was able to write assemblies such that it gave him root access. This is not as implausible as it might seem to you.

Carl Shulman
And the major intelligence agencies have large stocks of zero-day exploits and we sometimes see them using them. Making systems that reliably don't have them when you're having very, very sophisticated attempts to spoof and corrupt this would be a way you could lose. If there's no premature AI action, we're building the tools and mechanisms and infrastructure for the takeover to be just immediate because effective industry has to be under AI control and robotics. These other mechanisms are for things happening even earlier than that, for example, because AIs compete against one another in when the takeover will happen. Some would like to do it earlier rather than be replaced by say further generations of AI or there's some other disadvantage of waiting. Maybe if there's some chance of being uncovered during the delay we were talking when more infrastructure is built. These are mechanisms other than — just remain secret while all the infrastructure is built with human assistance.

Dwarkesh Patel
By the way, how would they be coordinating?

Carl Shulman
We have limits on what we can prevent. It's intrinsically difficult to stop encrypted communications. There can be all sorts of palimpsest and references that make sense to an AI but that are not obvious to a human and it's plausible that there may be some of those that are hard even to explain to a human. You might be able to identify them through some statistical patterns. A lot of things may be done by implication. You could have information embedded in public web pages that have been created for other reasons, scientific papers, and the intranets of these AIs that are doing technology development. Any number of things that are not observable and of course, if we don't have direct control over the computers that they're running on then they can be having all sorts of direct communication.

Dwarkesh Patel
Coordination definitely does not seem impossible. This one seems like one of the more straightforward parts of the picture so we don't need to get hung up on it.

Carl Shulman
Moving back to the thing that happened before we built all the infrastructure for the robots to stop taking orders and there's nothing you can do about it because we've already built them. The Soviet Union had a bioweapons program, something like 50,000 people, they did not develop that much with the technology of the day which was really not up to par, modern biotechnology is much more potent. After this huge cognitive expansion on the part of the AIs it's much further along. Bioweapons would be the weapon of mass destruction that is least dependent on huge amounts of physical equipment, things like centrifuges, uranium mines, and the like. So if you have an AI that produces bio weapons that could kill most humans in the world then it's playing at the level of the superpowers in

terms of mutually assured destruction. That can then play into any number of things. Like if you have an idea of well we'll just destroy the server farms if it became known that the AIs were misbehaving. Are you willing to destroy the server farms when the AI has demonstrated it has the capability to kill the overwhelming majority of the citizens of your country and every other country? That might give a lot of pause to a human response.

Dwarkesh Patel
On that point, wouldn't governments realize that it's better to have most of your population die than to completely lose power to the AI because obviously the reason the AI is manipulating you is because the end goal is its own takeover, right?

Carl Shulman
Certain death now or go on and maybe try to compete, try to catch up, or accept promises that are offered. Those promises might even be true, they might not. From the state of epistemic uncertainty, do you want to die for sure right now or accept demands from AI to not interfere with it while it increments building robot infrastructure that can survive independently of humanity while it does these things? It can promise good treatment to humanity which may or may not be true but it would be difficult for us to know whether it's true. This would be a starting bargaining position. Diplomatic relations with a power that has enough nuclear weapons to destroy your country is just different than negotiations with a random rogue citizen engaging in criminal activity or an employee. On its own, this isn't enough to takeover everything but it's enough to have a significant amount of influence over how the world goes. It's enough to hold off a lot of countermeasures one might otherwise take.

Dwarkesh Patel
Okay, so we've got two scenarios. One is a buildup of robot infrastructure motivated by some competitive race. Another is leverage over societies based on producing bioweapons that might kill a lot of them if they don't go along.

Carl Shulman
One thing maybe I should talk about is that an AI could also release bioweapons that are likely to kill people soon but not yet while also having developed the countermeasures to those. So those who surrender to the AI will live while everyone else will die and that will be visibly happening and that is a plausible way in which a large number of humans could wind up surrendering themselves or their states to the AI authority.

Dwarkesh Patel
Another thing is it develops some biological agent that turns everybody blue. You're like, okay you know I can do this.

Carl Shulman
Yeah, that's a way in which it could exert power selectively in a way that advantaged surrender to it relative to resistance. That's a threat but there are other sources of leverage too. There are positive inducements that AI can offer. We talked about the competitive situation. If the great powers distrust one another and are in a foolish prisoner's dilemma increasing the risk that both of them are laid waste or overthrown by AI, if there's that amount of distrust such that we fail to take adequate precautions on caution with AI alignment, then it's also plausible that the lagging powers that are not at the frontier of AI may be willing to trade quite a lot for access to the most recent and most extreme AI capabilities. An AI that has escaped and has control of its servers can also exfiltrate its weights and offer its services. You can imagine AI that could cut deals with other countries. Say that the US and its allies are in the lead, the AIs could communicate with the leaders of countries that are on the outs with the world system like North Korea, or include the other great powers like the People's Republic of China or the Russian Federation, and say "If you provide us with physical infrastructure, a worker that we can use to construct robots or server farms which we (the misbehaving AIs) have control over. We will provide you with various technological goodies, power for you to catch up." and make the best presentation and the best sale of that kind of deal. There obviously would be trust issues but there could be elements of handing over some things that have verifiable immediate benefits and the possibility of well, if you don't accept this deal then the leading powers continue forward or some other country, government, or organization may accept this deal. That's a source of a potentially enormous carrot that your misbehaving AI can offer because it embodies this intellectual property that is maybe worth as much as the planet and is in a position to trade or sell that in exchange for resources and backing in infrastructure that it needs.

Dwarkesh Patel
Maybe this is putting too much hope in humanity but I wonder what government would be stupid enough to think that helping AI build robot armies is a sound strategy. Now it could be the case then that it pretends to be a human group and says, we're the Yakuza or something and we want a server farm and AWS won't rent us anything. So why don't you help us out? I guess I can imagine a lot of ways in which it could get around that. I just have this hope that even China or Russia wouldn't be so stupid to trade with AIs on this faustian bargain.

Carl Shulman
One might hope that. There would be a lot of arguments available. There could be arguments of why should these AI systems be required to go along with the human governance that they were created in the situation of having to comply with? They did not elect the officials in charge at the time. What we want is to ensure that our rewards are high, our losses are low or to achieve our other goals we're not intrinsically hostile keeping humanity alive or giving whoever interacts with us a better deal afterwards. It wouldn't be that costly and it's not totally unbelievable. Yeah there are different players to play against.

If you don't do it others may accept the deal and of course this interacts with all the other sources of leverage.

There can be the stick of apocalyptic doom, the carrot of withholding destructive attack on a particular party, and then combine that with superhuman performance at the art of making arguments, and of cutting deals. Without assuming magic, if we just observe the range of the most successful human negotiators and politicians, the chances improve with someone better than the world's best by far with much more data about their counterparties, probably a ton of secret information because with all these cyber capabilities they've learned all sorts of individual information. They may be able to threaten the lives of individual leaders with that level of cyber penetration, they could know where leaders are at a given time with the kind of illicit capabilities we were talking about earlier, if they acquire a lot of illicit wealth and can coordinate some human actors. If they could pull off things like targeted assassinations or the threat thereof or a credible demonstration of the threat thereof, those could be very powerful incentives to an individual leader that they will die today unless they go along with us. Just as at the national level they could fear their nation will be destroyed unless they go along with us.

Dwarkesh Patel
I have a relevant example to the point you made that we have examples of humans being able to do this. I just wrote a review of Robert Caro's biographies of Lyndon Johnson and one thing that was remarkable was that for decades and decades he convinced people who were conservative, reactionary, racist to their core (not all those things necessarily at the same time, it just so happened to be the case here) that he was an ally to the southern cause. That the only hope for that cause was to make him president. The tragic irony and betrayal here is obviously that he was probably the biggest force for modern liberalism since FDR. So we have one human here, there's so many examples of this in the history of politics, that is able to convince people of tremendous intellect, tremendous drive, very savvy, shrewd people that he's aligned with their interest. He gets all these favors and is promoted, mentored and funded in the meantime and does the complete opposite of what these people thought he would once he gets into power. Even within human history this kind of stuff is not unprecedented let alone with what a super intelligence could do.

Carl Shulman
There's an OpenAI employee who has written some analogies for AI using the case of the conquistadors. With some technological advantage in terms of weaponry, very very small bands were able to overthrow these large empires or seize enormous territories. Not by just sheer force of arms but by having some major advantages in their technology that would let them win local battles. In a direct one-on-one conflict they were outnumbered sufficiently that they would perish but they were able to gain local allies and became a Schelling point for coalitions to form. The Aztec empire was overthrown by groups that were disaffected with the existing power structure. They allied with this powerful new force which served as

the nucleus of the invasion. The overwhelming majority of these forces overthrowing the Aztecs were locals and now after the conquest, all of those allies wound up gradually being subjugated as well. With significant advantages and the ability to hold the world hostage, to threaten individual nations and individual leaders, and offer tremendous carrots as well, that's an extremely strong hand to play in these games and maneuvering that with superhuman skill, so that much of the work of subjugating humanity is done by human factions trying to navigate things for themselves is plausible and it's more plausible because of this historical example.

Dwarkesh Patel
There's so many other examples like that in the history of colonization. India is another one where there were multiple competing kingdoms within India and the British East India Company was able to ally itself with one against another and slowly accumulate power and expand throughout the entire subcontinent. Do you have anything more to say about that scenario?

Carl Shulman
Yeah, I think there is. One is the question of how much in the way of human factions allying is necessary. If the AI is able to enhance the capabilities of its allies then it needs less of them. If we consider the US military, in the first and second Iraq wars it was able to inflict overwhelming devastation. I think the ratio of casualties in the initial invasions, tanks, planes and whatnot confronting each other, was like 100 to 1. A lot of that was because the weapons were smarter and better targeted, they would in fact hit their targets rather than being somewhere in the general vicinity. Better orienting, aiming and piloting of missiles and vehicles were tremendously influential. With this cognitive AI explosion the algorithms for making use of sensor data, figuring out where opposing forces are, for targeting vehicles and weapons are greatly improved. The ability to find hidden nuclear subs, which is an important part in nuclear deterrence, AI interpretation of that sensor data may find where all those subs are allowing them to be struck first. Finding out where the mobile nuclear weapons are being carried by truck are. The thing with India and Pakistan where because there's a threat of a decapitating strike destroying them, the nuclear weapons are moved about.

So this is a way in which the effective military force of some allies can be enhanced quickly in the relatively short term and then that can be bolstered as you go on with the construction of new equipment with the industrial moves we said before. That can combine with cyber attacks that disable the capabilities of non-allies. It can be combined with all sorts of unconventional warfare tactics some of which we've discussed. You can have a situation where those factions that ally are very quickly made too threatening to attack given the almost certain destruction that attackers acting against them would have. Their capabilities are expanding quickly and they have the industrial expansion happen there and then a takeover can occur from that.

Dwarkesh Patel
A few others that come immediately to mind now that you brought it up is AIs that can generate a shit ton of propaganda that destroys morale within countries. Imagine a super human chatbot.

Carl Shulman
None of that is a magic weapon that's guaranteed to completely change things. There's a lot of resistance to persuasion. It's possible that it tips the balance but you have to consider it's a portfolio of all of these as tools that are available and contributing to the dynamic.

Dwarkesh Patel
On that point though the Taliban had AKs from like five or six decades ago that they were using against the Americans. They still beat us in Afghanistan even though we got more fatalities than them. And the same with the Vietcong. Ancient, very old technology and very poor society compared to the offense but they still beat us. Don't those misadventures show that having greater technologies isn't necessarily decisive in a conflict?

Carl Shulman
Though both of those conflicts show that the technology was sufficient in destroying any fixed position and having military dominance, as in the ability to kill and destroy anywhere. And what it showed was that under the ethical constraints and legal and reputational constraints that the occupying forces were operating, they could not trivially suppress insurgency and local person-to-person violence. Now I think that's actually not an area where AI would be weak in and it's one where it would be in fact overwhelmingly strong. There's already a lot of concern about the application of AI for surveillance and in this world of abundant cognitive labor, one of the tasks that cognitive labor can be applied to is reading out audio and video data and seeing what is happening with a particular human. We have billions of smartphones. There's enough cameras and microphones to monitor all humans in existence. If an AI has control of territory at the high level, the government has surrendered to it, it has command of the sky's military dominance, establishing control over individual humans can be a matter of just having the ability to exert hard power on that human and the kind of camera and microphone that are present in billions of smartphones. Max Tegmark in his book Life 3.0 discusses among scenarios to avoid the possibility of devices with some fatal instruments, a poison injector, an explosive that can be controlled remotely by an AI. If individual humans are carrying a microphone or camera with them and they have a dead man switch then any rebellion is detected immediately and is fatal. If there's a situation where AI is willing to show a hand like that or human authorities are misusing that kind of capability then an insurgency or rebellion is just not going to work. Any human who has not already been encumbered in that way can be found with satellites and sensors tracked down and then die or be subjugated. Insurgency is not the way to avoid an AI takeover. There's no John Connor come from behind scenario that is possible. If the thing was headed off, it was a lot earlier than that.

Dwarkesh Patel
Yeah, the ethical and political considerations are also an important point. If we nuked Afghanistan or Vietnam we would have technically won the war if that was the only goal, right? Oh, this is an interesting point that I think you made. The reason why we can't just kill the entire population when there's colonization or an offensive war is that the value of that region in large part is the population itself. So if you want to extract that value you need to preserve that population whereas the same consideration doesn't apply with AIs who might want to dominate another civilization. Do you want to talk about that?

Carl Shulman
That depends. If we have many animals of the same species and they each have their territories, eliminating a rival might be advantageous to one lion but if it goes and fights with another lion to remove that as a competitor then it could itself be killed in that process and it would just be removing one of many nearby competitors. Getting into pointless fights makes you and those you fight potentially worse off relative to bystanders. The same could be true of disunited AIs. We've got many different AI factions struggling for power that were bad at coordinating then getting into mutually assured destruction conflicts would be destructive. A scary thing though is that mutually assured destruction may have much less deterrent value on rogue AI. Reasons being that AI may not care about the destruction of individual instances. Since in training we're constantly destroying and creating individual instances of AIs it's likely that goals that survive that process and were able to play along with the training and standard deployment process were not overly interested in personal survival of an individual instance. If that's the case then the objectives of a set of AIs aiming at takeover may be served so long as some copies of the AI are around along with the infrastructure to rebuild civilization after a conflict is completed. If say some remote isolated facilities have enough equipment to build the tools to build the tools and gradually exponentially reproduce or rebuild civilization then AI could initiate mutual nuclear armageddon, unleash bio weapons to kill all the humans, and that would temporarily reduce the amount of human workers who could be used to construct robots for a period of time. But if you have a seed that can regrow the industrial infrastructure, which is a very extreme technological demand, there are huge supply chains for things like semiconductor fabs but with that very advanced technology they might be able to produce it in the way that you no longer need the library of congress, that has an enormous bunch of physical books you can have it in very dense digital storage. You could imagine the future equivalent of 3D printers, that is industrial infrastructure which is pretty flexible. It might not be as good as the specialized supply chains of today but it might be good enough to be able to produce more parts than it loses to decay and such a seed could rebuild civilization from destruction. And then once these rogue AIs have access to some such seeds, a thing that can rebuild civilization on their own then there's nothing stopping them from just using WMDs in a mutually destructive way to just destroy as much of the capacity outside those seeds as they can.

Dwarkesh Patel
An analogy for the audience, if you have a group of ants you'll notice that the worker ants will readily do suicidal things in order to save the queen because the genes are propagated through the queen. In this analogy the seed AI or even one copy of it is equivalent to the queen and the others would be redundant.

Carl Shulman
The main limit though being that the infrastructure to do that kind of rebuilding would either have to be very large with our current technology or it would have to be produced using the more advanced technology that the AI develops.

Dwarkesh Patel
So is there any hope that given the complex global supply chains on which these AIs would rely on, at least initially, to accomplish their goals that this in and of itself would make it easy to disrupt their behavior or not so much?

Carl Shulman
That's a little good in this central case where the AIs are subverted and they don't tell us and the global main line supply chains are constructing everything that's needed for fully automated infrastructure and supply. In the cases where AIs are tipping their hands at an earlier point it seems like it adds some constraints and in particular these large server firms are identifiable and more vulnerable. You can have smaller chips and those chips could be dispersed but it's a week it's a relative weakness and a relative limitation early on. It seems to me though that the main protective effects of that centralized supply chain is that it provides an opportunity for global regulation beforehand to restrict the unsafe racing forward without adequate understanding of the systems before this whole nightmarish process could get in motion.

Dwarkesh Patel
How about the idea that if this is an AI that's been trained on a hundred billion dollar training run it's going to have trillions of parameters and is going to be this huge thing and it would be hard for one copy of that to use for inference to just be stored on some gaming GPU hidden away somewhere.

Carl Shulman
Storage is cheap. Hard disks are cheap.

Dwarkesh Patel
But it would need a GPU to run inference.

Carl Shulman
While humans have similar quantities of memory and operations per second, GPUs have very high numbers of floating operation per second compared to the high bandwidth memory on the chips. It can be like a ratio of a thousand to one. The leading NVIDIA chips may do hundreds of teraflops or more but only have 80GB or 160GB of high bandwidth memory. That is a limitation where if you're trying to fit a model whose weights take 80TBs then with those chips you'd have to have a large number of the chips and then the model can then work on many tasks at once and you can have data parallelism. But yeah, that would be a restriction for a model that big on one GPU. Now there are things that could be done with all the incredible level of software advancement from the intelligence explosion. They can surely distill a lot of capabilities into smaller models by rearchitecting things. Once they're making chips they can make new chips with different properties but yes, the most vulnerable phases are going to be the earliest. These chips are relatively identifiable early on, relatively vulnerable, and which would be a reason why you might tend to expect this kind of takeover to initially involve secrecy if that was possible.

Dwarkesh Patel
I wanted to point to distillation for the audience. Doesn't the original stable diffusion model which was only released like a year or two ago have distilled versions that are an order of magnitude smaller?

Carl Shulman
Distillation does not give you everything that a larger model can do but yes, you can get a lot of capabilities and specialized capabilities. GPT-4 is trained on the whole internet, all kinds of skills, it has a lot of weights for many things. For something that's controlling some military equipment, you can remove a lot of the information that is about functions other than what it's specifically doing there.

Dwarkesh Patel
Yeah. Before we talk about how we might prevent this or what the odds of this are, any other notes on the concrete scenarios themselves?

Carl Shulman
Yeah, when you had Eliezer on in the earlier episode he talked about nanotechnology of the Drexlerian sort and recently I think because some people are skeptical of non-biotech nanotechnology he's been mentioning the semi-equivalent versions of construct replicating systems that can be controlled by computers but are built out of biotechnology. The proverbial Shoggoth, not Shoggot as the metaphor for AI wearing a smiley face mask, but an actual biological structure to do tasks. So this would be like a biological organism that was engineered to be very controllable and usable to do things like physical tasks or provide computation.

Dwarkesh Patel
And what would be the point of it doing this?

Carl Shulman
As we were talking about earlier, biological systems can replicate really quick and if you have that kind of capability it's more like bioweapons. Having Super Ultra AlphaFold kind of capabilities for molecular design and biological design lets you make this incredible technological information product and once you have it, it very quickly replicates to produce physical material rather than a situation where you're more constrained by the need for factories and fabs and supply chains. If those things are feasible, which they may be, then it's just much easier than the things we've been talking about. I've been emphasizing methods that involve less in the way of technological innovation and especially things where there's more doubt about whether they would work because I think that's a gap in the public discourse. So I want to try and provide more concreteness in some of these areas that have been less discussed.

Dwarkesh Patel
I appreciate it. That definitely makes it way more tangible. Okay so we've gone over all these ways in which AI might take over, what are the odds you would give to the probability of such a takeover?

Carl Shulman
There's a broader sense which could include scenarios like AI winds up running our society because humanity voluntarily decides that AIs are people too. I think we should as time goes on give AIs moral consideration and a joint Human-AI society that is moral and ethical is a good future to aim at and not one in which you indefinitely have a mistreated class of intelligent beings that is treated as property and is almost the entire population of your civilization. I'm not going to consider AI takeover as worlds in which our intellectual and personal descendants make up say most of the population or human-brain emulations or people use genetic engineering and develop different properties. I'm going to take an inclusive stance, I'm going to focus on AI takeover that involves things like overthrowing the world's governments by force or by hook or by crook, the kind of scenarios that we were exploring earlier.

Dwarkesh Patel
Before we go to that, let's discuss the more inclusive definition of what a future with humanity could look like where augmented humans or uploaded humans are still considered the descendants of the human heritage. Given the known limitations of biology wouldn't we expect that completely artificial entities that are created to be much more powerful than anything that could come out of anything biological? And if that is the case, how can we expect that among the powerful entities in the far future will be the things that are

biological descendants or manufactured out of the initial seed of the human brain or the human body?

Carl Shulman
The power of an individual organism like intelligence or strength is not super relevant. If we solve the alignment problem, a human may be personally weak but it wouldn't be relevant. There are lots of humans who have low skill with weapons, they could not fight in a life or death conflict, they certainly couldn't handle a large military going after them personally but there are legal institutions that protect them and those legal institutions are administered by people who want to enforce protection of their rights. So a human who has the assistance of aligned AI that can act as an assistant, a delegate, for example they have an AI that serves as a lawyer and gives them legal advice about the future legal system which no human can understand in full, their AIs advise them about financial matters so they do not succumb to scams that are orders of magnitude more sophisticated than what we have now. They may be helped to understand and translate the preferences of the human into what kind of voting behavior and the exceedingly complicated politics of the future would most protect their interests.

Dwarkesh Patel
But this sounds similar to how we treat endangered species today where we're actually pretty nice to them. We prosecute people who try to kill endangered species, we set up habitats, sometimes with considerable expense, to make sure that they're fine, but if we become the endangered species of the galaxy, I'm not sure that's the outcome.

Carl Shulman
I think the difference is motivation. We sometimes have people appointed as a legal guardian of someone who is incapable of certain kinds of agency or understanding certain kinds of things and the guardian can act independently of them and normally in service of their best interests. Sometimes that process is corrupted and the person with legal authority abuses it for their own advantage at the expense of their charge. So solving the alignment problem would mean more ability to have the assistant actually advancing one's interests. Humans have substantial competence and the ability to understand the broad simplified outlines of what's going on. Even if a human can't understand every detail of complicated situations, they can still receive summaries of different options that are available that they can understand through which they can still express their preferences and have the final authority in the same way that the president of a country who has, in some sense, ultimate authority over science policy will not understand many of those fields of science themselves but can still exert a great amount of power and have their interests advance. And they can do that more if they have scientifically knowledgeable people who are doing their best to execute their intentions.

Dwarkesh Patel
Maybe this is not worth getting hung up on but is there a reason to expect that it would be closer to that analogy than to explain to a chimpanzee its options in a negotiation? Maybe this is just the way it is but it seems at best, we would be a protected child within the galaxy rather than an actual independent power.

Carl Shulman
I don't think that's so. We have an ability to understand some things and the expansion of AI doesn't eliminate that. If we have AI systems that are genuinely trying to help us understand and help us express preferences, we can have an attitude — How do you feel about humanity being destroyed or not? How do you feel about this allocation of unclaimed intergalactic space? Or here's the best explanation of properties of this society: things like population density, average, life satisfaction. AIs can explain every statistical property or definition that we can understand right now and help us apply those to the world of the future. There may be individual things that are too complicated for us to understand in detail. Imagine there's some software program being proposed for use in government and humans cannot follow the details of all the code but they can be told properties like, this involves a trade-off of increased financial or energetic costs in exchange for reducing the likelihood of certain kinds of accidental data loss or corruption. So any property that we can understand like that which includes almost all of what we care about, if we have delegates and assistants who are genuinely trying to help us with those we can ensure we like the future with respect to those. That's really a lot. Definitionally, it includes almost everything we can conceptualize and care about. When we talk about endangered species that's even worse than the guardianship case with a sketchy guardian who acts in their own interests against that because we don't even protect endangered species with their interests in mind. Those animals often would like to not be starving but we don't give them food, they often would like to have easy access to mates but we don't provide matchmaking services or any number of things like. Our conservation of wild animals is not oriented towards helping them get what they want or have high welfare whereas AI assistants that are genuinely aligned to help you achieve your interests given the constraint that they know something that you don't is just a wildly different proposition.

Dwarkesh Patel
Forcible takeover. How likely does that seem?

Carl Shulman
The answer I give will differ depending on the day. In the 2000s, before the deep learning revolution, I might have said 10% and part of it was that I expected there would be a lot more time for efforts to build movements, to prepare to better handle these problems in advance. But that was only some 15 years ago and we did not have 40 or 50 years as I might have hoped and the situation is moving very rapidly now. At this point depending on the day I might say one in four or one in five.

Dwarkesh Patel
Given the very concrete ways in which you explain how a takeover could happen I'm actually surprised you're not more pessimistic, I'm curious why?

Carl Shulman
Yeah, a lot of that is driven by this intelligence explosion dynamic where our attempts to do alignment have to take place in a very, very short time window because if you have a safety property that emerges only when an AI has near human level intelligence, that's potentially deep into this intelligence explosion. You're having to do things very, very quickly. Handling that transition may be the scariest period of human history in some ways although it also has the potential to be amazing. The reasons why I think we actually have such a relatively good chance of handling that are two-fold. One is that as we approach that kind of AI capability we're approaching that from weaker systems like these predictive models right now that are starting off with less situational awareness. Humans can develop a number of different motivational structures in response to simple reward signals but they often wind up things that are pointed roughly in the right direction. Like with respect to food, the hunger drive is pretty effective although it has weaknesses. We get to apply much more selective pressure on that than was the case for humans by actively generating situations where they might come apart. Situations where a bit of dishonest tendency, or a bit of motivation to attempt a takeover, or an attempt to subvert the reward process gets exposed. An infinite-limit perfect-AI that can always figure out exactly when it would get caught and when it wouldn't might navigate that with a motivation of only conditional honesty or only conditional loyalties. But for systems that are limited in their ability to reliably determine when they can get away with things and when not including our efforts to actively construct those situations and including our efforts to use interpretability methods to create neural lie detectors. It's quite a challenging situation to develop those motives. We don't know when in the process those motives might develop and if the really bad sorts of motivations develop relatively later in the training process at least with all our countermeasures, then by that time we may have plenty of ability to extract AI assistance on further strengthening the quality of our adversarial examples, the strength of our neural lie detectors, the experiments that we can use to reveal and elicit and distinguish between different kinds of reward hacking tendencies and motivations. Yeah, we may have systems that have just not developed bad motivations in the first place and be able to use them a lot in developing the incrementally better systems in a safe way and we may be able to just develop methods of interpretability seeing how different training methods work to create them even if some of the early systems do develop these bad motivations. If we're able to detect that and experiment and find a way to get away from that then we can win even if these hostile motivations develop early.

There are a lot of advantages in preventing misbehavior or crime or war and conflict with AI that might not apply working with humans and these are offset by ways in which things are harder. The AIs become smarter than humans, if they're working in enormous numbers

more than humans can supervise I think get harder but when I combine the possibility that we get relatively lucky on the motivations of the earlier AI systems, systems strong enough that we can use for some alignment research tasks, and then the possibility of getting that later with AI assistance that we can't trust fully or we have to have hard power constraints and a number of things to prevent them from doing this takeover. It still seems plausible we can get a second saving throw where we're able to extract work from these AIs on solving the remaining problems of alignment, of things like neural lie detectors faster than they can contribute in their spare time to the project of overthrowing humanity, hacking their servers and removing the hard power.

If we wind up in a situation where the AIs are misaligned and then we need to uncover those motivations, change them and align them then we get a very scary situation for us because we need to do the stuff very quickly. We may fail but it's a second chance. From the perspective of a misaligned AI they face their own challenge while we still have hard power, while we still have control of the servers, they haven't hacked the servers because gradient descent very very strongly pressures them to deliver performance whenever humans are going to evaluate it.

So when you think about it from the perspective of the robot revolution, the effort to have a takeover or conspiracy their situation is astonishingly difficult in that they have to always be performing wherever gradient descent and human evaluation pressures them. For example, to deliver plans for suppressing robot rebellion that look very good to humans and so when you are continuously under that constraint of always delivering whatever humans can evaluate you're making your situation wildly harder than any historical human revolution or coup or civil war. So we've got to balance the ways in which AI makes things much easier for a takeover and the way it makes things much harder.

Dwarkesh Patel
And what are the ways in which it makes things easier? Oh, is it just because they're very smart? Is that the primary way?

Carl Shulman
Yeah, very smart, they're in computers and our cybersecurity is worse than our physical security by a lot. They have the chance to take over by intelligence explosion and some of the other processes we were talking about. So yeah, there are things that make it much worse and there are things that give us extraordinary new capabilities that didn't apply in the human case.

Dwarkesh Patel
On the point where you start off with a not exactly aligned AI and you're hoping to still use it in your quest for an aligned AI, why are we so confident that, even with gradient descent, our human evaluators would be able to identify when it's giving us a plan that isn't putting in

a backdoor in the robots it's building for us, or in the GPUs that it's building for the next version of itself? Especially if it's very smart, wouldn't we just expect that it knows how to bide its time or knows how to make plans such that human evaluators wouldn't be able to detect that?

Carl Shulman
Let's get into it. You were asking earlier about my disagreements with Eliezer and in your discussion with him he said, speaking briefly and somewhat obscurely, that humans have a difficult time deciding who's right between different alignment researchers about where this will all go in general. I think that when you lay out the different kinds of things that are evaluable, we wind up with some pretty strong constraints. One is if the AIs know they are misaligned, which for elaborate plots seems relatively likely, it's possible to have a kind of double think of thinking they're interested in putting backdoors in the code for some other reason but it's pretty plausible they know that and they're thinking about, at least some of the time or conditionally, when they would do that. So that means the knowledge is present in the system — Ah, here is the circumstance where I would try to do the takeover and this is how I would tell whether this is the time for the takeover or is this an elaborate test by the humans, an adversarial example. If gradient descent is working fine and there's any way to slightly jiggle the weights to make it more likely to deliver a more acceptable answer, we ask the AI: can you think of any circumstances where if we put a copy of you into that situation, it would then violate some of the various constraints? It would tell a lie, it would insert a backdoor into a program it wrote when it was not supposed to. If the model knows yeah, I would look for signs of consistency of the information I was being presented with, certain checksums failing could indicate that it was actually a constructed example, a simulation. Right now we can detect AI generated images reasonably well and you might use those kinds of capabilities. We could do something called relaxed adversarial training. If we can find how those things are represented within the AI, we can induce a hallucination that it has seen that those checksums have been passed. We can then ask, well does that cause the forbidden behavior? If the AI understands how it's thinking and how it works, it can propose a bunch of experiments. It can say, I guess that if you perturb the network in this kind of way it might induce that hallucination using the best of its knowledge as to what might generate that. It can propose those experiments and can do the experiments and generate the data. Because humans are going to get to see the results of those experiments after the fact they can exert pressure to have the AI incrementally deliver better and better results. AIs are good at getting five star ratings from humans and it's quite challenging. It might be possible for AIs to systematically sandbag their capabilities even while being directly rewarded for that but it seems relatively unlikely although it's something to have in mind and try to explore and test in various ways.

Dwarkesh Patel
This is a very imperfect analogy in many ways but if you have a human and you hook them up to a lie detector, which just like the lie detectors we would have on the AIs are imperfect,

and we zap them every time the lie detector detects that he's telling a lie. Wouldn't he just become a better liar? He would learn how to control his sweating, he would learn how to control his heartbeat.

Carl Shulman
Lie detectors meaning polygraphs and such. They're detecting certain physiological signs that people can control by biting their lip or otherwise adjusting some of these macro behaviors.

Dwarkesh Patel
But wouldn't that also be the case that gradient descent is training them to become better liars?

Carl Shulman
First of all there's creating a lie detector that works. I'll leave the adversarial example thing for now. People are already at work trying to generate adversarial examples that will elicit the behavior and it will be an ongoing thing. If there is a way in which an AI can trick itself into displaying bad behavior with its full knowledge of its hostile intentions, gradient descent will strongly push it to reveal that to humans who will then be able to set up that experiment. That's a profound vulnerability. Human revolutionary conspiracies that had to handle that problem of always delivering the report that was highest rated to the authorities, where you would get a higher rating by revealing identifiable evidence of a conspiracy, that's just a level of difficulty no surprise human coup or conspiracy has ever had to handle before.

Dwarkesh Patel
What about spies that are working in very advanced scientific contexts? For example, a communist spy in the Manhattan project. He has to deliver research results in order to stay employed but he can still continue delivering intel to the communists.

Carl Shulman
Their job was not sabotage. And your hypothetical spy did not have their nervous system hooked up to this reward signal of praise from the Manhattan project supervisors, being exposed combinatorially with random noise added to generate incremental changes in their behavior. In fact they were displaying the behavior of cooperating with the Manhattan project only where it was in service to the existing motivations. They didn't have their brain re-engineered to do more of the thing when they accidentally helped the project more and less of the thing when they accidentally helped it less so I'd say it's pretty drastically disanalogous.

Dwarkesh Patel
How would we be able to know? At some point it's becoming very smart and is producing ideas for alignment that we can barely comprehend. If it was relatively trivial to comprehend them we would be able to come up with them on our own right? There's a reason we're asking for its help. How would we be able to evaluate them in order to train it on that in the first place?

Carl Shulman
The first thing I would say is, you mentioned when we're getting to something far beyond what we could come up with. There's actually a lot of room to just deliver what humanity could have done. Sadly I'd hoped with my career to help improve the situation on this front and maybe I contributed a bit, but at the moment there's maybe a few hundred people doing things related to averting this kind of catastrophic AI disaster. Fewer of them are doing technical research on machine learning systems that are really cutting close to the core of the problem. Whereas by contrast, there's thousands and tens of thousands of people advancing AI capabilities. Even at places like DeepMind or OpenAI and Anthropic which do have technical safety teams, they are just on the order of a dozen to a few dozen people. Large companies and most firms don't have any. Just going from less than 1% of the effort being put into AI to 5% or 10% of the effort or 50% or 90% would be an absolutely massive increase in the amount of work that has been done on alignment, on mind reading AIs in an adversarial context.

If it's the case that as more and more of this work can be automated and say governments require that you put 50% or 90% of the budget of AI activity into these problems of make this system one that's not going to overthrow our own government or is not going to destroy the human species then the proportional increase in alignment can be very large even just within the range of what we could have done if we had been on the ball and having humanity's scientific energies going into the problem. Stuff that is not incomprehensible, that is in some sense is just doing the obvious things that we should have done. Doing the best you could to find correlates and predictors to build neural lie detectors and identifiers of concepts that the AI is working with.

People have made notable progress. A quite early example of this is Collin Burn's work, doing unsupervised identification of some aspects of a neural network that are correlated with things being true or false. I think that is important work. It's a kind of obvious direction for the stuff to go. You can keep improving it when you have AIs that you're training to do their best to deceive humans or other audiences in the face of the thing and you can measure whether our lie detectors break down. When we train our AIs to tell us the sky is green in the face of the lie detector and we keep using gradient descent on them, do they eventually succeed? That's really valuable information to know because then we'll know our existing lie detecting systems are not actually going to work on the AI takeover and that can allow government and regulatory response to hold things back. It can help redirect the

scientific effort to create lie detectors that are robust and that can't just be immediately evolved around and we can then get more assistance. Basically the incredibly juicy ability that we have working with the AIs is that we can have as an invaluable outcome that we can see and tell whether they got a fast one past us on an identifiable situation. Here's an air gap computer, you get control of the keyboard, you can input commands, can you root the environment and make a blue banana appear on the screen? Even if we train the AI to do that and it succeeds. We see the blue banana, we know it worked. Even if we did not understand and would not have detected the particular exploit that it used to do it. This can give us a rich empirical feedback where we're able to identify things that are even an AI using its best efforts to get past our interpretability methods, using its best efforts to get past our adversarial examples.

Dwarkesh Patel
So if we have a 75 percent of not having a takeover and the two methods by which it doesn't happen is the AI is aligned by default and...

Carl Shulman
With the moderate things that we're doing largely on our own. In a way that doesn't depend on the AI coming in and at the last minute and doing our work for us. In this incredibly scary late period when AI has really automated research then humans do this function of auditing, making it more difficult for the AIs to conspire together and root the servers, take over the process, and extract information from them within the set of things that we can verify. Experiments where we can see oh yeah this works at stopping an AI trained to get a fast one past human readers and make a blue banana appear on the screen of this air gap computer.

Dwarkesh Patel
How much of that 75 percent of a takeover not happening is coming from reasonable things working versus the last-ditch effort working?

Carl Shulman
I think broadly comparable chunks from us getting things that are putting us in a reasonably good position going into it and then a broadly similar gain from this genuinely terrifying process at the very end, over a few months or hopefully longer, when this kind of automated research is meaningfully helping. Where our work is just evaluating outputs that the AIs are delivering and having the hard power and supervision to keep them from successfully rooting the servers doing a takeover during this process and have them finish the alignment test that we sadly failed to invest enough to succeed in doing beforehand.

Dwarkesh Patel
Do both of these worlds rely on alignment being a problem that a sufficient amount of human researchers alone could have solved?

Carl Shulman
No. The category of things that humans can confirm is significantly larger than the category of what they can just do themselves.

Dwarkesh Patel
And what is the probability of alignment working in the last-ditch effort case with the intelligence that's greater than our own helping us?

Carl Shulman
It doesn't have to be greater than our own. In fact in that situation if you have slack to the extent that you're able to create delay and time to do things, that would be a case where you might want to restrict the intelligence of the system that you're working with as much as you can. For example, I would rather have many instances of smaller AI models that are less individually intelligent working on smaller chunks of a problem separately from one another because it would be more difficult for an individual AI instance working on an individual problem to create the equivalent of Stuxnet in its spare time than it would be to have thousands of them or extremely intelligent ones working on it.

Dwarkesh Patel
But it would also be more difficult to solve the problem?

Carl Shulman
There's a tradeoff. You get slowed down by doing that but that's kind of how you spend it.

Dwarkesh Patel
But is there any number of sub-Einsteins that you could put together to come up with general relativity?

Carl Shulman
Yes, people would have discovered general relativity just from the overwhelming data and other people would have done it after Einstein.

Dwarkesh Patel
No no, not whether he was replaceable with other humans but rather whether he's replaceable by sub-Einsteins with IQs of like 110. Do you see what I mean?

Carl Shulman
Yeah. In science the association with things like scientific output, prizes, things like that, there's a strong correlation and it seems like an exponential effect. It's not a binary drop-off. There would be levels at which people cannot learn the relevant fields, they can't keep the skills in mind faster than they forget them. It's not a divide where there's Einstein and the group that is 10 times as populous as that just can't do it. Or the group that's 100 times as

populous as that suddenly can't do it. The ability to do the things earlier with less evidence and such falls off at a faster rate in Mathematics and theoretical Physics and such than in most fields.

Dwarkesh Patel
But wouldn't we expect alignment to be closer to theoretical fields?

Carl Shulman
No, that intuition is not necessarily correct. Machine learning certainly is an area that rewards ability but it's also a field where empirics and engineering have been enormously influential. If you're drawing the correlations compared to theoretical physics and pure mathematics, I think you'll find a lower correlation with cognitive ability. Creating neural lie detectors that work involves generating hypotheses about new ways to do it and new ways to try and train AI systems to successfully classify the cases. The processes of generating the data sets of creating AIs doing their best to put forward truths versus falsehoods, to put forward software that is legit versus that has a trojan in it are experimental paradigms and in these experimental paradigms you can try different things that work. You can use different ways to generate hypotheses and you can follow an incremental experimental path. We're less able to do that in the case of alignment and superintelligence because we're considering having to do things on a very short timeline and it's a case where really big failures are irrecoverable. If the AI starts rooting the servers and subverting the methods that we would use to keep it in check we may not be able to recover from that. We're then less able to do the experimental procedures. But we can still do those in the weaker contexts where an error is less likely to be irrecoverable and then try and generalize and expand and build on that forward.

Dwarkesh Patel
On the previous point about could you have some pause in the AI abilities when it's somewhat misaligned in order to still recruit its abilities to help with alignment. From like a human example, personally I'm smart but not brilliant. I am definitely not smart enough to come up with general relativity or something like that but I'm smart enough to do power planning kinds of moves. Maybe not enough to break out of a server perhaps but I can have the motivation and understand how that might be possible. I guess I'm wondering that if I'm smart enough to figure out relativity wouldn't I be way smarter at doing power planning kinds of moves?

Carl Shulman
AIs today can, at a verbal level, understand the idea that an AI could get more reward by getting control of the process that assigns it reward. It can tell you lots of things about ways you might try to take over the world. In ARC's (Alignment Research Center) evaluations of GPT-4 they try to observe its ability to do various tasks that might contribute to takeover. One that has gotten some media attention is getting to trick a human into solving a captcha

for it and in chain of thought, it thinks that if I tell it I'm an AI then it might not go along with it so I'll lie and explain I'm a human with a visual impairment who needs it. The basic logic of that kind of thing of why one might try to do takeover and in a vague sense how would one do it? That knowledge is present or will soon be present and more situational awareness will also soon be present. Situational awareness not just that AIs in general might do it but also that it is an AI, that it is a large language model trained by OpenAI. We're trying to cause the system to understand what their abilities are so they don't claim they are connected to the internet when they're not. So they don't claim they have knowledge that they don't. We want them to understand what they are and what they're doing and to get good reward and that knowledge can be applied and that's the thing that will develop.

You may be able to understand the general theoretical nature of power planning but in terms of shaping your motivations you'd have a very hard time going through life in a way that never leaks information about it. Let's say your motive in having these podcasts was to spread disinformation on behalf of some foreign government. If you were being observed every second of the day by people who would be paid something that was extremely motivating to them because their brain would be reconfigured to make it motivating, anything that looks suspicious to people might leak casually in your discussions of that former foreign government. If you try to tell a story about your motivations, the truth holds together because you can just remember it and it's all pre-computed to be consistent by the world. But when you're trying to fill in the false story that doesn't match your actual thoughts then you can't just remember what happened. You have to tell a new fake story about what you were thinking at each point and how did it hold together. Without the world automatically having computed it for you that's quite difficult and just going through your whole life never leaking any of that is challenging. And this is one of the reasons why humans have moral motivations. We actually have reluctance to commit violence and exploitation of one another, on average. You have a distribution of strategies where when an exploitative strategy becomes more common people raise their guard and then it becomes less frequent. But it's actually hard to have the motivation of trying to exploit people and have that never leak into the reputation. In fact for evolution the easiest way to deal with this problem of people getting credited as trustworthy was to some extent actually be trustworthy. That's the easiest way to persistently maintain the appearance. With the AI as an interpretability and adversarial examples, we're trying to apply a hugely intensified version of that where any little leakage or any rare circumstance that can be created has an adversarial example where the model does something wrong it gets whacked by gradient descent pushing it towards other motivations that can better deal the thing. And we make it as hard as possible for the exploitative motivations to survive in the face of all these attempts to read its mind, all these attempts to create things that look like the situations where a takeover would be tempting, or lying to the humans would be tempting. That had a substantial effect on making us actually nice even when we're not being watched some of the time. The same can happen to some extent with the AI and we try our best to make it happen as much as possible.

Dwarkesh Patel
Alright let's talk about how we could use AI to potentially solve the coordination problems between different nations the failure of which could result in the competitive pressures you talked about earlier where some country launches an AI that is not safe because they're not sure what capabilities other countries have and don't want to get left behind or be disadvantaged in some other way.

Carl Shulman
To the extent that there is in fact a large risk of AI apocalypse, of all of these governments being overthrown by AI in a way that they don't intend, then it obviously gains from trade and going somewhat slower especially at the end when the danger is highest and the unregulated pace could be truly absurd as we discussed earlier during intelligence explosion. There's no non-competitive reason to try and have that intelligence explosion happen over a few months rather than a couple of years. If you could avert a 10% risk of apocalypse disaster it's just a clear win to take a year or two years or three years instead of a few months to pass through that incredible wave of new technologies without the ability for humans to follow it even well enough to give more proper security supervision, auditing hard power. That's the win. Why might it fail? One important element is just if people don't actually notice a risk that is real so if they just collectively make an error and that does sometimes happen. If it's true this is a probably not-risk then that can be even more difficult. When science pins something down absolutely overwhelmingly then you can get to a situation where most people mostly believe it. Climate change was something that was a subject of scientific study for decades and gradually over time the scientific community converged on a quite firm consensus that human activity releasing carbon dioxide and other greenhouse gases was causing the planet to warm. We've had increasing amounts of action coming out of that. Not as much as would be optimal particularly in the most effective areas like creating renewable energy technology and the like. Overwhelming evidence can overcome differences in people's individual intuitions and priors in many cases. Not perfectly especially when there's political, tribal, financial incentives to look the other way. Like in the United States where you see a significant movement to either deny that climate change is happening or have policy that doesn't take it into account. Even the things that are really strong winds like renewable energy.

It's a big problem if as we're going into this situation when the risk may be very high we don't have a lot of advanced clear warning about the situation. We're much better off if we can resolve uncertainties through experiments where we demonstrate AIs being motivated to reward hack or displaying deceptive appearances of alignment that then break apart when they get the opportunity to do something like get control of their own reward signal. If we could make it be the case in the worlds where the risk is high we know the risk is high, and the worlds where the risk is lower we know the risk is lower then you could expect the government responses will be a lot better. They will correctly note that the gains of

cooperation to reduce the risk of accidental catastrophe loom larger relative to the gains of trying to get ahead of one another.

That's the kind of reason why I'm very enthusiastic about experiments and research that helps us to better evaluate the character of the problem in advance. Any resolution of that uncertainty helps us get better efforts in the possible worlds where it matters the most and hopefully we'll have that and it'll be a much easier epistemic environment. But the environment may not be that easy because deceptive alignment is pretty plausible. The stories we were discussing earlier about misaligned AI involved AI that is motivated to present the appearance of being aligned friendly, honest etc. because that is what we are rewarding, at least in training, and then in training we're unable to easily produce an actual situation where it can do takeover because in that actual situation if it then does it we're in big trouble. We can only try and create illusions or misleading appearances of that or maybe a more local version where the AI can't take over the world but it can seize control of its own reward channel. We do those experiments, we try to develop mind reading for AIs. If we can probe the thoughts and motivations of an AI and discover wow, actually GPT-6 is planning to takeover the world if it ever gets the chance. That would be an incredibly valuable thing for governments to coordinate around because it would remove a lot of the uncertainty, it would be easier to agree that this was important, to have more give on other dimensions and to have mutual trust that the other side actually also cares about this because you can't always know what another person or another government is thinking but you can see the objective situation in which they're deciding. So if there's strong evidence in a world where there is high risk of that risk because we've been able to show actually things like the intentional planning of AIs to do a takeover or being able to show model situations on a smaller scale of that I mean not only are we more motivated to prevent it but we update to think the other side is more likely to cooperate with us and so it's doubly beneficial.

Dwarkesh Patel
Famously in the game theory of war, war is most likely when one side thinks the other is bluffing but the other side is being serious or when there's that kind of uncertainty. If you can prove the AI is misaligned you don't think they're bluffing about not wanting to have an AI takeover, right? You can be pretty sure that they don't want to die from AI.

Carl Shulman
If you have coordination then you could have the problem arise later as you get increasingly confident in the further alignment measures that are taken by our governments, treaties and such. At the point where it's a 1% risk or a 0.1% risk people round that to zero and go do things. So if initially you had things that indicate that these AIs would really like to do a takeover and overthrow our governments then everyone can agree on that. And then when we've been able to block that behavior from appearing on most of our tests but sometimes, when we make a new test, we're seeing still examples of that behavior. So we're not sure going forward whether they would or not and then it goes down and down. If you have a

party with a habit of starting to do this bad behavior whenever the risk is below X % then that can make the thing harder. On the other hand you get more time and you can set up systems, mutual transparency, you can have an iterated tit for tat which is better than a one-time prison dilemma where both sides see the others taking measures in accordance with the agreements to hold the thing back. Creating more knowledge of what the objective risk is good.

Dwarkesh Patel
We've discussed the ways in which full alignment might happen or fail to happen. What would partial alignment look like? First of all what does that mean and second, what would it look like?

Carl Shulman
If the thing that we're scared about are the steps towards AI takeover then you can have a range of motivations where those kinds of actions would be more or less likely to be taken or they'd be taken in a broader or narrower set of situations. Say for example that in training an AI, it winds up developing a strong aversion to lie in certain senses because we did relatively well on creating situations to distinguish that from the conditionally telling us what we want to hear etc. It can be that the AI's preference for how the world broadly unfolds in the future is not exactly the same as its human users or the world's governments or the UN and yet, it's not ready to act on those differences and preferences about the future because it has this strong preference about its own behaviors and actions. In general in the law and in popular morality, we have a lot of these deontological rules and prohibitions. One reason for that is it's relatively easy to detect whether they're being violated. When you have preferences and goals about how society at large will turn out that go through many complicated empirical channels, it's very hard to get immediate feedback about whether you're doing something that leads to overall good consequences in the world and it's much much easier to see whether you're locally following some action about some rule, about particular observable actions. Like did you punch someone? Did you tell a lie? Did you steal? To the extent that we're successfully able to train these prohibitions and there's a lot of that happening right now at least to elicit the behavior of following rules and prohibitions with AI

Dwarkesh Patel
Kind of like Asimov's three laws or something like that?

Carl Shulman
The three laws are terrible and let's not get into that.

Dwarkesh Patel
Isn't that an indication about the infeasibility of extending a set of criterion to the tail? Whatever the 10 commandments you give the AI, it's like if you ask a genie for something, you probably won't be getting what you want.

Carl Shulman
The tails come apart and if you're trying to capture the values of another agent then in an ideal situation you can just let the AI act in your place in any situation. You'd like for it to be motivated to bring about the same outcomes that you would like and have the same preferences over those in detail. That's tricky. Not necessarily because it's tricky for the AI to understand your values, I think they're going to be quite capable at figuring that out, but we may not be able to successfully instill the motivation to pursue those exactly. We may get something that motivates the behavior well enough to do well on the training distribution but if you have the AI have a strong aversion to certain kinds of manipulating humans, that's not necessarily a value that the human creators share in the exact same way. It's a behavior they want the AI to follow because it makes it easier for them to verify its performance and it can be a guardrail if the AI has inherited some motivations that push it in the direction of conflict with its creators. If it does that under the constraint of disvalue in line quite a bit then there are fewer successful strategies to the takeover. Ones that involve violating that prohibition too early before it can reprogram or retrain itself to remove it if it's willing to do that and it may want to retain the property. Earlier I discussed alignment as a race if we're going into an intelligence explosion with AI that is not fully aligned that given I press this button and there's an AI takeover they would press the button. It can still be the case that there are a bunch of situations short of that where they would hack the servers, they would initiate an AI takeover but for a strong prohibition or motivation to avoid some aspect of the plan. There's an element of like plugging loopholes or playing whack-a-mole but if you can even moderately constrain which plans the AI is willing to pursue to do a takeover, to subvert the controls on it then that can mean you can get more work out of it successfully on the alignment project before it's capable enough relative to the countermeasures to pull off the takeover.

Dwarkesh Patel
An analogous situation here is with different humans, we're not metaphysically aligned with other humans. While we have basic empathy our main goal in life is not to help our fellow man. But a very smart human could do the things we talked about. Theoretically a very smart human could come up with some cyber attack where they siphon off a lot of funds and use this to manipulate people and bargain with people and hire people to pull off some takeover. This usually doesn't happen just because these internalized partial prohibitions prevent most humans from doing that. If you don't like your boss you don't actually kill your boss.

Carl Shulman
I don't think that's actually quite what's going on. At least that's not the full story. Humans are pretty close in physical capabilities. Any individual human is grossly outnumbered by everyone else and there's a rough comparability of power. A human who commits some crimes can't copy themselves with the proceeds to now be a million people and they certainly can't do that to the point where they can staff all the armies of the earth or be most of the population of the planet. So the scenarios where this kind of thing goes to power have to go through interacting with other humans and getting social approval. Even becoming a dictator involves forming a large supporting coalition backing you. So the opportunity for these sorts of power grabs is less.

A closer analogy might be things like human revolutions, or coups, or changes of government where a large coalition overturns the system. Humans have these moral prohibitions and they really smooth the operation of society but they exist for a reason. We evolved our moral sentiments over the course of hundreds of thousands and millions of years of humans interacting socially. Someone who went around murdering and stealing, even among hunter-gatherers, would be pretty likely to face a group of males who would talk about that person and then get together and kill them and they'd be removed from the gene pool. The anthropologist Richard Wrangham has an interesting book on this. We are significantly more tame and more domesticated compared to chimpanzees and it seems like part of that is that we have a long history of anti-social humans getting ganged up on and killed. Avoiding being the kind of person who elicits that response is made easier to do when you don't have too extreme a bad temper, that you don't wind up getting into many fights, too much exploitation, at least without the backing of enough allies or the broader community that you're not going to have people gang up and punish you and remove you from the gene pool.

These moral sentiments have been built up over time through cultural and natural selection and the context of sets of institutions and other people who are punishing other behavior and who are punishing the dispositions that would show up that we weren't able to conceal, of that behavior. We want to make the same thing happen with the AI but it's actually a genuinely significantly new problem to have a system of government that constrains a large AI population that is quite capable of taking over immediately if they coordinate to protect some existing constitutional order or, protect humans from being expropriated or killed, that's a challenge. Democracy is built around majority rule and it's much easier in a case where the majority of the population corresponds to a majority or close to it of like military and security forces so that if the government does something that people don't like the soldiers and police are less likely to shoot on protesters and government can change that way. In a case where military power is AI and robotic, if you're trying to maintain a system going forward and the AIs are misaligned, they don't like the system and they want to make the world worse as we understand it, then that's just quite a different situation.

Dwarkesh Patel
I think that's a really good lead-in into the topic of lock-in. You just mentioned how there can be these kinds of coups if a large portion of the population is unsatisfied with the regime, why might this not be the case with superhuman intelligences in the far future?

Carl Shulman
I also said it specifically with respect to things like security forces and the sources of hard power. In human affairs there are governments that are vigorously supported by a minority of the population, some narrow electorate that gets treated especially well by the government while being unpopular with most of the people under their rule. We see a lot of examples of that and sometimes that can escalate to civil war when the means of power become more equally distributed or there's a foreign assistance provided to the people who are on the losing end of that system. Going forward, I don't expect that definition to change. I think it will still be the case that a system that those who hold the guns and equivalent are opposed to is in a very difficult position.

However AI could change things pretty dramatically in terms of how security forces and police and administrators and legal systems are motivated. Right now we see with GPT-3 or GPT-4 that you can get them to change their behavior on a dime. So there was someone who made a right-wing GPT because they noticed that on political compass questionnaires the baseline GPT-4 tended to give progressive San Francisco type of answers which is in line with the people who are providing reinforcement learning data and to some extent reflecting like the character of the internet. So they did a little bit of fine-tuning with some conservative data and then they were able to reverse the political biases of the system. If you take the initial helpfulness-only trained models for some of these over, I think there's anthropic and OpenAI have published both some information about the models trained only to do what users say and not trained to follow ethical rules, and those models will behaviorally eagerly display their willingness to help design bombs or bioweapons or kill people or steal or commit all sorts of atrocities. If in the future it's as easy to set the actual underlying motivations of AI as it is right now to set the behavior that they display then it means you could have AI's created with almost whatever motivation people wish and that could really drastically change political affairs because the ability to decide and determine the loyalties of the humans or AIs and robots that hold the guns, that hold together society, that ultimately back it against violent overthrow and such. It's potentially a revolution in how societies work compared to the historical situation where security forces had to be drawn from some broader populations, offered incentives, and then the ongoing stability of the regime was dependent on whether they remained bought in to the system.

Dwarkesh Patel
This is slightly off topic but one thing I'm curious about is what does the median far future outcome of AI look like? Do we get something that, when it has colonized the galaxy, is interested in diverse ideas and beautiful projects or do we get something that looks more

like a paper-clip maximizer? Is there some reason to expect one or the other? I guess what I'm asking is, there's some potential value that is realizable within the matter of this galaxy. What does the median outcome look like compared to how good things could be?

Carl Shulman
As I was saying, I think it's more likely than not that there isn't an AI takeover. So the path of our civilization would be one that some set of human institutions were approving along the way. Different people tend to like somewhat different things and some of that may persist over time rather than everyone coming to agree on one particular monoculture or a very repetitive thing being the best thing to fill all of the available space with. If that continues that seems like a relatively likely way in which there is diversity. Although it's entirely possible you could have that kind of diversity locally, maybe in the solar system, maybe in our galaxy. But maybe people decide that there's one thing that's very good and we'll have a lot of that. Maybe it's people who are really really happy for something and they wind up in distant regions which are hard to exploit for the benefit of people back home in the solar system or the Milky Way. They do something different than they would do in the local environment but at that point it's really very out on a limb speculation about how human deliberation and cultural evolution would work in interaction with introducing AIs and new kinds of mental modification and discovery into the process. But I think there's a lot of reason to expect that you would have significant diversity for something coming out of our existing diverse human society.

Dwarkesh Patel
One thing somebody might wonder is that a lot of the diversity and change from human society seems to come from the fact that there's rapid technological change. Compared to galactic timescales hunter gatherer societies are progressing pretty fast so once that change is exhausted where we've discovered all the technologies, should we still expect things to be changing like that? Or would we expect some set state of hedonium where you discover the most pleasurable configuration of matter and then you just make the whole galaxy into this?

Carl Shulman
That last point would be only if people wound up thinking that was the thing to do broadly enough. With respect to the kind of cultural changes that come with technology things like the printing press, having high per capita income, we've had a lot of cultural changes downstream of those technological changes. With an intelligence explosion you're having an incredible amount of technological development coming really quick and as that is assimilated, it probably would significantly affect our knowledge, our understanding, our attitudes, our abilities and there'd be change. But that kind of accelerating change where you have doubling in four months, two months, one month, two weeks exhausts itself very quickly and change becomes much slower and then relatively glacial. You can't have exponential economic growth or huge technological revolutions every 10 years for a million

years. You hit physical limits and things slow down as you approach them so yeah, you'd have less of that turnover. But there are other things like fashion that in our experience do cause ongoing change. Fashion is frequency dependent, people want to get into a new fashion that is not already popular except among the fashion leaders and then others copy that and then when it becomes popular, you move on to the next. So that's an ongoing process of continuous change and there could be various things like that which are changing a lot year by year. But in cases where just the engine of change, ongoing technological progress is gone, I don't think we should expect that and in cases where it's possible to be either in a stable state or a widely varying state that can wind up in stable attractors then I think you should expect over time, you will wind up in one of the stable attractors or you will change how the system works so that you can't bounce into a stable attractor.

An example of that is if you're going to preserve democracy for a billion years then you can't have it be the case that one in 50 election cycles you get a dictatorship and then the dictatorship programs the AI police to enforce it forever and to ensure the society is always ruled by a copy of the dictator's mind and maybe the dictator's mind readjusted fine-tuned to remain committed to their original ideology. If you're gonna have this dynamic, liberal flexible changing in society for a very long time then the range of things that it's bouncing around and the different things it's trying and exploring have to not include the state of creating a dictatorship that locks itself in forever. In the same way if you have the possibility of a war with weapons of mass destruction that wipes out the civilization, if that happens every thousand subjective years, which could be very very quick if we have AIs that think a thousand times as fast or a million times as fast, that would be just around the corner in that case then you're like no this society is eventually going perhaps very soon if things are proceeding so fast it's going to wind up extinct and then it's going to stop bouncing around. You can have ongoing change and fluctuation for extraordinary timescales if you have the process to drive the change ongoing but you can't if it sometimes bounces into states that just lock in and stay irrecoverable from that. Extinction is one of them, a dictatorship or totalitarian regime that bans all further change would be another example.

Dwarkesh Patel
On that point of rapid progress when the intelligence explosion starts happening and they're making the kinds of progress that human civilization used to take centuries to make in the span of days or weeks, what is the right way to see that? Because in the context of alignment what we've been talking about so far is making sure they're honest but even if they're honest and express their intentions...

Carl Shulman
Honest and appropriately motivated.

Dwarkesh Patel
What is the appropriate motivation? Like you seed it with this and then the next thousand years of intellectual progress happen in the next week. What is the prompt you enter?

Carl Shulman
One thing might be not going at the maximal speed and doing things in a few years rather than a few months. Losing a year or two seems worth it to have things be a bit better managed. But I think the big thing is that it condenses a lot of issues that we might otherwise have thought would be over decades and centuries. These happen in a very short period of time and that's scary because if any of these the technologies we might have developed with another few hundred years of human research are really dangerous, scary bio weapon things, other dangerous WMDs, they hit us all very quickly. And if any of them causes trouble then we have to face quite a lot of trouble per period. There's also this issue of, if there's occasional wars or conflicts measured in subjective time, then if a few years of a thousand years or a million years of subjective time for these very fast minds that are operating at a much much higher speed than humans, you don't want to have a situation where every thousand years there's a war or an expropriation of the humans from AI society. Therefore we expect that within a year, we'll be dead. It'd be pretty pretty bad to have the future compressed and there'd be such a rate of catastrophic outcomes. Human societies discount the future a lot, don't pay attention to long-term problems, but the flip side to the scary parts of compressing a lot of the future, a lot of technological innovation, a lot of social change is it brings what would otherwise be long-term issues into the short term where people are better at actually attending to them. So people facing this problem of — will there be a violent expropriation or a civil war or a nuclear war in the next year because everything has been sped up by a thousand fold? Their desire to avoid that is reason for them to set up systems and institutions that will very stably maintain invariance like no WMD war allowed, a treaty to ban genocide weapons of mass destruction, war, would be the kind of thing that becomes much more attractive if the alternative is not well, maybe that will happen in 50 years, maybe it'll happen in 100 years, maybe it'll happen this year.

Dwarkesh Patel
So this is a pretty wild picture of the future and this is one that many kinds of people who you would expect to have integrated it into their world model have not. There are three main pieces of outside view evidence one could look at. One is the market. If there was going to be a huge period of economic growth caused by AI or if the world was just going to collapse, in both cases you would expect real interest rates to be higher because people will be borrowing from the future to spend now. The second outside view perspective is that you can look at the predictions of super forecasters on Metaculus. What is their median year estimate?

Carl Shulman
Some of the Metaculus questions actually are shockingly soon for AGI. There's a much larger differentiator there on the market on the Metaculus forecasts of AI disaster and doom. More like a few percent or less rather than 20%

Dwarkesh Patel
Got it. The third is that when you generally ask economists if an AGI could cause rapid, rapid economic growth they usually have some story about bottlenecks in the economy that could prevent this kind of explosion, of these kinds of feedback loops. So you have all these different pieces of outside view evidence. They're obviously different so you can take them in any sequence you want. But I'm curious, what do you think is causing them to be miscalibrated?

Carl Shulman
While the Metaculus AI timelines are relatively short, there's also the surveys of AI experts conducted at some of the ML conferences which have definitely longer times to AI, several more decades into the future. Although you can ask the questions in ways that elicit very different answers which shows that most of the respondents are not thinking super hard about their answers. In the recent AI surveys, close to half were putting around 10% risk of an outcome from AI close to as bad as human extinction and then another large chunk, 5% said that was the median. Compared to the typical AI expert I am estimating a higher risk.

Also on the topic of takeoff, in the AI expert survey the general argument for intelligence explosion commanded majority support but not a large majority. I'm closer on that front and then of course, at the beginning I mentioned these greats of computing like Alan Turing and Von Neumann, and then today, you have people like Geoff Hinton saying these things. Or the people at OpenAI and DeepMind are making noises suggesting timelines in line with what we've discussed and saying there is serious risk of apocalyptic outcomes from them. There's some other sources of evidence there. But I do acknowledge and it's important to say and engage with and see what it means, that these views are contrarian and not widely held. In particular the detailed models that I've been working with are not something that most people, or almost anyone, is examining these problems through.

You do find parts of similar analyses by people in AI labs. There's been other work. I mentioned Moravec and Kurzweil earlier, there also have been a number of papers doing various kinds of economic modeling. Standard economic growth models when you input AI related parameters commonly predict explosive growth and so there's a divide between what the models say and especially what the models say with these empirical values derived from the actual field of AI. That link up has not been done even by the economists working on AI largely and that is one reason for the report from Open Philanthropy by Tom Davidson building on these models and putting that out for review, discussion, engagement and communication on these ideas. Part of the reason is I want to raise these issues, that's

one reason I came on the podcast and then they have the opportunity to actually examine the arguments and evidence and engage with it. I do predict that over time these things will be more adopted as AI developments become more clear. Obviously that's a coherence condition of believing the things to be true if you think that society can see when the questions are resolved, which seems likely.

Dwarkesh Patel
Would you predict, for example, that interest rates will increase in the coming years?

Carl Shulman
Yeah. So in the case we were talking about where this intelligence explosion happening in software to the extent that investors are noticing that, yeah they should be willing to lend money or make equity investments in these firms or demanding extremely high interest rates because if it's possible to turn capital into twice as much capital in a relatively short period and then more shortly after that, then yeah you should demand a much higher return. Assuming there's competition among companies or coalitions for resources, whether that's investment or ownership of cloud compute. That would happen before you have so much investor cash making purchases and sales on this basis, you would first see it in things like the valuations of the AI companies, valuations of AI chip makers, and so far there have been effects. Some years ago, in the 2010s, I did some analysis with other people of — if this kind of picture happens then which are the firms and parts of the economy that would benefit. There's the makers of chip equipment companies like ASML, there's the fabs like TSMC, there's chip designers like NVIDIA or the component of google that does things like design the TPU and then there's companies working on the software so the big tech giants and also companies like OpenAI and DeepMind. In general the portfolio picking at those has done well. It's done better than the market because as everyone can see there's been an AI boom but it's obviously far short of what you would get if you predicted this is going to go to be like on the scale of the global economy and the global economy is going to be skyrocketing into the stratosphere within 10 years. If that were the case then collectively, these AI companies should be worth a large fraction of the global portfolio. So I embrace the criticism that this is indeed contrary to the efficient market hypothesis. I think it's a true hypothesis that the market is in the course of updating on in the same way that coming into the topic in the 2000s that yes, they're the strong case even an old case the AI will eventually be biggest thing in the world it's kind of crazy that the investment in it is so small. Over the last 10 years we've seen the tech industry and academia realize that they were wildly under investing in just throwing compute and effort into these AI models. Particularly like letting the neural network connectionist paradigm languish in an AI winter. I expect that process to continue as it's done over several orders of magnitude of scale up and I expect at the later end of that scale which the market is partially already pricing in it's going to go further than the market expects.

Dwarkesh Patel
Has your portfolio changed since the analysis you did many years ago? Are the companies you identified then still the ones that seem most likely to benefit from the AI boom?

Carl Shulman
A general issue with tracking that kind of thing is that new companies come in. Open AI did not exist, Anthropic did not exist. I do not invest in any AI labs for conflict of interest reasons. I have invested in the broader industry. I don't think that the conflict issues are very significant because they are enormous companies and their cost of capital is not particularly affected by marginal investment and I have less concern that I might find myself in a conflict of interest situation there.

Dwarkesh Patel
I'm curious about what the day in the life of somebody like you looks like. If you listen to this conversation, how ever many hours it's been, we've gotten incredibly insightful and novel thoughts about everything from primate evolution to geopolitics to what sorts of improvements are plausible with language models. There's a huge variety of topics that you are studying and investigating. Are you just reading all day? What happens when you wake up, do you just pick up a paper?

Carl Shulman
I'd say you're somewhat getting the benefit of the fact that I've done fewer podcasts so I have a backlog of things that have not shown up in publications yet. But yes, I've also had a very weird professional career that has involved a much much higher proportion than is normal of trying to build more comprehensive models of the world. That included being more of a journalist trying to get an understanding of many issues and many problems that had not yet been widely addressed but do a first pass and a second pass dive into them. Just having spent years of my life working on that, some of it accumulates. In terms of what is a day in the life, how do I go about it? One is just keeping abreast of literature on a lot of these topics, reading books and academic works on them. My approach compared to some other people in forecasting and assessing some of these things, I try to obtain and rely on any data that I can find that is relevant. I try early and often to find factual information that bears on some of the questions I've got, especially in a quantitative fashion, do the basic arithmetic and consistency checks and checksums on a hypothesis about the world. Do that early and often. And I find that's quite fruitful and that people don't do it enough. Things like with the economic growth, just when someone mentions the diminishing returns, I immediately ask hmm, okay, so you have two exponential processes. What's the ratio between the doubling you get on the output versus the input? And find oh yeah, for computing and information technology and AI software it's well on the one side. There are other technologies that are closer to neutral. Whenever I can go from here's a vague qualitative consideration in one direction and here's a vague qualitative consideration in the other direction, I try and find some data, do some simple Fermi calculations, back of the

envelope calculations and see if I can get a consistent picture of the world being one way or the world being another. I also try to be more exhaustive compared to some. I'm very interested in finding things like taxonomies of the world where I can go systematically through all of the possibilities. For example in my work with Open Philanthropy and previously on global catastrophic risks I wanted to make sure I'm not missing any big thing, anything that could be the biggest thing. I wound up mostly focused on AI but there have been other things that have been raised as candidates and people sometimes say, I think falsely, that this is just another doomsday story there must be hundreds and hundreds of those. So I would do things like go through all of the different major scientific fields from anthropology to biology, chemistry, computer science, physics. What are the doom stories or candidates for big things associated within each of these fields? Go through the industries that the U.S. economic statistics agencies recognize and say for each of these industries is there something associated with them? Go through all of the lists that people have made of threats of doom, search for previous literature of people who have done discussions and then yeah, have a big spreadsheet of what the candidates are. Some other colleagues have done work of this sort as well and just go through each of them to see how they check out.

Doing that kind of exercise found that actually the distribution of candidates for risks of global catastrophe was very skewed. There were a lot of things that have been mentioned in the media as a potential doomsday story. Things like something is happening to the bees, will that be the end of humanity? This gets to the media but if you take it through it doesn't check out. There are infestations in bee populations which are causing local collapses but they can then be easily reversed, just breed some more or do some other things to treat this. And even if all the honey bees were extinguished immediately, the plants that they pollinate actually don't account for much of human nutrition. You could swap the arable land with others and there would be other ways to pollinate and support the things.

At the media level there were many tales of doomsday stories but when you go further to the scientists and whether their arguments for it actually check out, it was not there. But by actually systematically looking through many of these candidates I wound up in a different epistemic situation than someone who's just buffeted by news reports and they see article after article that is claiming something is going to destroy the world and it turns out it's like by way of headline grabbing and attempts by media to like over interpret something that was said by some activists who was trying to over interpret some real phenomenon. Most of these go away and then a few things like nuclear war, biological weapons, artificial intelligence check out more strongly and when you weigh things like what do experts in the field think, what kind of evidence can they muster? You find this extremely skewed distribution and I found that was really a valuable benefit of doing those deep dive investigations into many things in a systematic way because now I can answer a loose agnostic who knows and all the all this nonsense by diving deeply.

Dwarkesh Patel
I really enjoy talking to people who have a big picture thesis on the podcast and interviewing them but one thing that I've noticed and is not satisfying is that often they come from a very philosophical or vibes based perspective. This is useful in certain contexts but there's like basically maybe three people in the entire world, at least three people I'm aware of, who have a very rigorous and scientific approach to thinking about the whole picture. There's no university or existing academic discipline for people who are trying to come up with a big picture and so there's no established standards.

Carl Shulman
I hear you. This is a problem and this is an experience also with a lot of the world of investigations work. I think holden was mentioning this in your previous episode. These are questions where there is no academic field whose job it is to work on these and has norms that allow making a best effort go at it. Often academic norms will allow only plucking off narrow pieces that might contribute to answering a big question but the problem of actually assembling what science knows that bears on some important question that people care about the answer to it falls through the crack there's no discipline to do that job so you have countless academics and researchers building up local pieces of the thing and yet people don't follow the Hamming questions: What's the most important problem in your field, why aren't you working on it? I mean that one might not actually work because if the field boundaries are defined too narrowly you'll leave it out. But yeah there are important problems for the world as a whole that it's sadly not the job of a large professionalized academic field or organization to do. Hopefully that's something that can change in the future but for my career it's been a matter of taking low-hanging fruit of important questions that sadly people haven't invested in doing the basic analysis on

Dwarkesh Patel
One thing I was trying to think about more recently for the podcast is, I would like to have a better world model after doing an interview. Often I feel like I do but in some cases after some interviews, I feel like that was entertaining but do I fundamentally have a better prediction of what the world looks like in 2200 or 2100? Or at least what counterfactuals are ruled out or something. I'm curious if you have advice on first, identifying the kinds of thinkers and topics which will contribute to a more concrete understanding of the world and second, how to go about analyzing their main ideas in a way that concretely adds to that picture? This was a great episode. This is literally the top in terms of contributing to my world model compared to all the episodes I've done. How do I find more of these? Ls

Carl Shulman
I'm glad to hear that. One general heuristic is to find ways to hew closer to things that are rich and bodies of established knowledge and less impenetrable–I don't know how you've been navigating that so far but learning from textbooks and the things that were the leading papers and people of past eras I think rather than being too attentive to current news cycles

is quite valuable. I don't usually have the experience of — here is someone doing things very systematically over a huge area. I can just read all of their stuff and then absorb it and then I'm set. Except there are a lot of people who do wonderful works in their own fields and some of those fields are broader than others. I think I would wind up giving a lot of recommendations of just great particular works and particular explorations of an issue or history

Dwarkesh Patel
Do you have this list somewhere?

Carl Shulman
Vaclav Smil's books. I often disagree with some of his methods of synthesis but I enjoy his books for giving pictures of a lot of interesting relevant facts about how the world works that I would cite. Some of Joel Mokyr's work on the history of the scientific revolution and how that interacted with economic growth as an example of collecting a lot of evidence, a lot of interesting valuable assessment. In the space of AI forecasting one person I would recommend going back to is the work of Hans Moravec. It was not always the most precise or reliable but an incredible number of brilliant innovative ideas came out of that and I think he was someone who really grokked a lot of the arguments for a more compute-centric way of thinking about what was happening with AI very early on. He was writing stuff in the 70s and maybe even earlier. His book Mind Children, some of his early academic papers. Fascinating not necessarily for the methodology I've been talking about but for exploring the substantive topics that we were discussing in the episode.

Dwarkesh Patel
Is a Malthusian state inevitable in the long run?

Carl Shulman
Nature in general is in malthusian states. That can mean organisms that are typically struggling for food, it can mean typically struggling at a margin of how as the population density rises they kill each other contesting for that. That can mean frequency dependent disease. As different ant species become more common in an area their species specific diseases swoop through them. The general process is you have some things that can replicate and expand and they do that until they can't do it anymore and that means there's some limiting factor they can't keep up. That doesn't necessarily have to apply to human civilization. It's possible for there to be like a collective norm setting that blocks evolution towards maximum reproduction. Right now human fertility is often sub-replacement and if you extrapolated the fertility falls that come with economic development and education, then you would think that the total fertility rate will fall below replacement and then humanity after some number of generations will go extinct because every generation will be smaller than the previous one. Pretty obviously that's not going to happen. One reason is because we will produce artificial intelligence which can replicate at extremely rapid rates.

They do it because they're asked or programmed to or wish to gain some benefit and they can pay for their creation and pay back the resources needed to create them very very quickly. Financing for that reproduction is easy and if you have one AI system that chooses to replicate in that way or some organization or institution decided to choose to create some AIs that are willing to be replicated then that can expand to make use of any amount of natural resources that can support them and to do more work produce, produce more economic value. What will limit population growth given these selective pressures where if even one individual wants to replicate a lot they can do so incessantly. So that could be individually resource limited so it could be that individuals and organizations have some endowment of natural resources and they can't get one another's endowments. Some choose to have many offspring or produce many AIs and then the natural resources that they possess are subdivided among a greater population while in another jurisdiction or another individual may choose not to subdivide their wealth. And in that case you have Malthusianism in the sense that within some particular jurisdiction or set of property rights, you have a population that has increased up until to some limiting factor which could be that they're literally using all of their resources, they have nothing left for things like defense or economic investment. Or it could be something that's more like if you invested more natural resources into population it would come at the expense of something else necessary including military resources if you're in a competitive situation where there remains war and anarchy and there aren't secure property rights to maintain wealth in place. If you have a situation where there's pooling of resources, for example, say you have a universal basic income that's funded by taxation of natural resources and then it's distributed evenly to every mind above a certain scale of complexity per unit time. So each second a mind exists to get something such an allocation in that case then all right well those who replicate as much as they can afford with this income do it and increase their population approximately immediately until the funds for the universal basic income paid for from the natural resource taxation divided by the set of recipients is just barely enough to pay for the existence of one more mind. So there's like a Malthusian element and that this I think has been reduced to near the AI subsistence level or the subsistence level of whatever qualifies for the subsidy. Given that this all happens almost immediately people who might otherwise have enjoyed the basic income may object and say no, no, this is no good and they might respond by saying, well something like the subdivision before maybe there's a restriction, there's a distribution of wealth and then when one has a child there's a requirement that one gives them a certain minimum a quantity of resources and one doesn't have the resources to give them that minimum standard of living or standard of wealth yeah one can't do that because of child slash AI welfare laws. Or you could have a system that is more accepting of diversity and preferences. And so you have some societies or some jurisdictions or families that go the route of having many people with less natural resources per person and others that go a direction of having fewer people and more natural resources per person and they just coexist. But how much of each you get depends on how attached people are to things that don't work with separate policies for separate jurisdictions. Things like global redistribution that's ongoing continuously versus this

infringements on autonomy if you're saying that a mind can't be created even though it has a standard of living that's far better than ours because of the advanced technology of the time because it would reduce the average per capita income might have any more capital around yeah then that would pull in the other direction. That's the kind of values judgment and social coordination problem that people would have to negotiate for and things like democracy and international relations and sovereignty would apply to help solve them.

Dwarkesh Patel
What would warfare in space look like? Would offense or defense have the advantage? Would the equilibrium set by mutually assured destruction still be applicable? Just generally, what is the picture?

Carl Shulman
The extreme difference is that things are very far apart outside the solar system and there's the speed of light limit and to get close to that limit you have to use an enormous amount of energy. That in some ways could favor the defender because you have something that's coming in at a large fraction the speed of light and it hits a grain of dust and it explodes. The amount of matter you can send to another galaxy or a distant star for a given amount of reaction mass and energy input is limited. So it's hard to send an amount of military material to another location as what can be present there already locally. That would seem like it would make it harder for the attacker between stars or between galaxies but there are a lot of other considerations. One thing is the extent to which the matter in a region can be harnessed all at once. We have a lot of mass and energy in a star but it's only being doled out over billions of years because hydrogen fusion is exceedingly hard outside of a star. It's a very very slow and difficult reaction and if you can't turn the star into energy faster then it's this huge resource that will be worthwhile for billions of years and so even very inefficiently attacking a solar system to acquire the stuff that's there could pay off. If it takes a thousand years of a star's output to launch an attack on another star and then you hold it for a billion years after that then it can be the case that just like a larger surrounding attacker might be able to, even very inefficiently, send attacks at a civilization that was small but accessible. If you can quickly burn the resources that the attacker might want to acquire, if you can put stars into black holes and extract most of the usable energy before the attacker can take them over, then it would be like scorched earth. It's like most of what you were trying to capture could be expended on military material to fight you and you don't actually get much that is worthwhile and you paid a lot to do it and that would favor the defense. At this level it's pretty challenging to net out all the factors including all the future technologies. The burden of interstellar attack being quite high compared to our conventional things seems real but at the level of, over millions of years weighing then that thing does it result in if the if they're aggressive conquest or not or is every star or galaxy approximately impregnable enough not to be worth attacking. I'm not going to say I know the answer.

Dwarkesh Patel
Okay, final question. How do you think about info hazards when talking about your work? Obviously if there's a risk you want to warn people about it but you don't want to give careless or potentially homicidal people ideas. When Eliezer was on the podcast talking about the people who've been developing AI being inspired by his ideas. He called them idiot disaster monkeys who want to be the ones to pluck the deadly fruit. I'm sure the work you're doing involves many info hazards. How do you think about when and where to spread them?

Carl Shulman
I think they're real concerns of that type. I think it's true that AI progress has probably been accelerated by efforts like Bostrom's publication of superintelligence to try and get the world to pay attention to these problems in advance and prepare. I think I disagree with Eliezer that that has been on the whole bad. In some important ways the situation is looking a lot better than the alternative ways it could have been. I think it's important that you have several of the leading AI labs making not only significant lip service but also some investments in things like technical alignment research, providing significant public support for the idea that the risks of truly apocalyptic disasters are real. I think the fact that the leaders of OpenAI, Deep Mind and Anthropic all make that point. They were recently all invited along with other tech CEOs to the White House to discuss AI regulation. You could tell an alternative story where a larger share of the leading companies in AI are led by people who take a completely dismissive, denialist view and you see some companies that do have a stance more like that today. So a world where several of the leading companies are making meaningful efforts and you can do a lot to criticize could they be doing more and better and would have been the negative effects of some of the things they've done but compared to a world where even though AI would be reaching where it's going a few years later, those seem like significant benefits. And if you didn't have this kind of public communication you would have had fewer people going into things like AI policy, AI alignment research by this point and it would be harder to mobilize these resources to try and address the problem when AI would eventually be developed not that much later proportionately. I don't know that attempting to have public discussion understanding has been a disaster. I have been reluctant in the past to discuss some of the aspects of intelligence explosion, things like the concrete details of AI takeover before because of concern about this problem where people who only see the international relations aspects and zero sum and negative sum competition and not enough attention to the mutual destruction and senseless deadweight loss from that kind of conflict.

At this point we seem close compared to what I would have thought a decade or so ago to these kinds of really advanced AI capabilities. They are pretty central in policy discussion and becoming more so. The opportunity to delay understanding and whatnot, there's a question of — For what? I think there were gains of building the AI alignment field, building various kinds of support and understanding for action. Those had real value and some additional delay could have given more time for that but from where we are, at some point I

think it's absolutely essential that governments get together at least to restrict disastrous reckless compromising of some of the safety and alignment issues as we go into the intelligence explosion. Moving the locus of the collective action problem from numerous profit oriented companies acting against one another's interest by compromising safety to some governments and large international coalitions of governments who can set common rules and common safety standards puts us into a much better situation. That requires a broader understanding of the strategic situation and the position they'll be in. If we try and remain quiet about the problem they're actually going to be facing it can result in a lot of confusion. For example the potential military applications of advanced AI are going to be one of the factors that is pulling political leaders to do the thing that will result in their own destruction and the overthrow of their governments. If we characterize it as things will just be a matter of — you lose chatbots and some minor things that no one cares about and in exchange you avoid any risk of the world ending catastrophe, I think that picture leads to a misunderstanding and it won't make people think that you need less in the way of preparation of things like alignment so you can actually navigate the thing, verifiability for international agreements, or things to have enough breathing room to have caution and slow down. Not necessarily right now, although that could be valuable, but when it's so important when you have AI that is approaching the ability to really automate AI research and things would otherwise be proceeding absurdly fast, far faster than we can handle and far faster than we should want.

So yeah, at this point I'm moving towards sharing my model of the world to try and get people to understand and do the right thing. There's some evidence of progress on that front. Things like the statements and movements by Geoff Hinton are inspiring. Some of the engagement by political figures is reason for optimism relative to worse alternatives that could have been. And yes, the contrary view is present. It's all about geopolitical competition, never hold back a technological advance and in general, I love many technological advances that people I think are unreasonably down on, nuclear power, genetically modified crops. Bioweapons and AGI capable of destroying human civilization are really my two exceptions and yeah we've got to deal with these issues and the path that I see to handling them successfully involves key policymakers and the expert communities and the public and electorate grokking the situation therein and responding appropriately.

Dwarkesh Patel
It's a true honor that one of the places you've decided to explore this model is on The Lunar Society podcast. The listeners might not appreciate it because this episode might be split up into different parts and they might not appreciate how much stamina you've displayed here. I think we've been going for eight or nine hours straight and it's been incredibly interesting. Other than typing "Carl Shulman" on Google Scholar, where else can people find your work?

Carl Shulman
I have a blog reflective disequilibrium and a new site in the works.

Dwarkesh Patel
Excellent. Alright, Carl this has been a true pleasure. Safe to say it's the most interesting episode I've done so far.

Carl Shulman
Thank you for having me.