Dwarkesh Podcast #55 - Carl Shulman (Pt 1) - Intelligence Explosion, Primate Evolution, Robot Doublings, & Alignment

Published - June 15, 2023

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I have the pleasure of speaking with Carl Shulman. Many of my former guests, and this is not an exaggeration, have told me that a lot of their biggest ideas have come directly from Carl especially when it has to do with the intelligence explosion and its impacts. So I decided to go directly to the source and we have Carl today on the podcast. He keeps a super low profile but is one of the most interesting intellectuals I've ever encountered and this is actually his second podcast ever. We're going to go deep into the heart of many of the most important ideas that are circulating right now directly from the source. Carl is also an advisor to the Open Philanthropy project which is one of the biggest funders on causes having to do with AI and its risks, not to mention global health and well being. And he is a research associate at the Future of Humanity Institute at Oxford. So Carl, it's a huge pleasure to have you on the podcast. Thanks for coming.

Carl Shulman
Thank you Dwarkesh. I've enjoyed seeing some of your episodes recently and I'm glad to be on the show.

Dwarkesh Patel
Excellent, let's talk about AI. Before we get into the details, give me the big picture explanation of the feedback loops and just general dynamics that would start when you have something that is approaching human-level intelligence.

Carl Shulman
The way to think about it is — we have a process now where humans are developing new computer chips, new software, running larger training runs, and it takes a lot of work to keep Moore's law chugging (while it was, it's slowing down now). And it takes a lot of work to develop things like transformers, to develop a lot of the improvements to AI neural networks. The core method that I want to highlight on this podcast, and which I think is underappreciated, is the idea of input-output curves. We can look at the increasing difficulty of improving chips and sure, each time you double the performance of computers it’s harder and as we approach physical limits eventually it becomes impossible. But how much harder? There's a paper called “Are Ideas Getting Harder to Find?" that was published a few years ago. 10 years ago at MIRI, I did an early version of this analysis using data mainly from Intel and the large semiconductor fabricators. In this paper they cover a period where the productivity of computing went up a million fold, so you could get a million times the computing operations per second per dollar, a big change but it got harder. The amount of investment and the labor force required to make those continuing advancements went up and up and up. It went up 18 fold over that period. Some take this to say — “Oh, diminishing returns. Things are just getting harder and harder and so that will be the end of progress eventually.” However in a world where AI is doing the work, that doubling of computing performance, translates pretty directly to a doubling or better of the effective labor supply. That is, if when we had that million-fold compute increase we used it to run artificial

intelligences who would replace human scientists and engineers, then the 18x increase in the labor demands of the industry would be trivial. We're getting more than one doubling of the effective labor supply than we need for each doubling of the labor requirement and in that data set, it's over four. So when we double compute we need somewhat more researchers but a lot less than twice as many. We use up some of those doublings of compute on the increasing difficulty of further research, but most of them are left to expedite the process. So if you double your labor force, that's enough to get several doublings of compute. You use up one of them on meeting the increased demands from diminishing returns. The others can be used to accelerate the process so you have your first doubling take however many months, your next doubling can take a smaller fraction of that, the next doubling less and so on. At least in so far as the outputs you're generating, compute for AI in this story, are able to serve the function of the necessary inputs. If there are other inputs that you need eventually those become a bottleneck and you wind up more restricted on this.

Dwarkesh Patel
Got it. The bloom paper said there was a 35% increase in transistor density and there was a 7% increase per year in the number of researchers required to sustain that pace.

Carl Shulman
Something in the vicinity, yeah. Four to five doublings of compute per doubling of labor inputs.

Dwarkesh Patel
I guess there's a lot of questions you can delve into in terms of whether you would expect a similar scale with AI and whether it makes sense to think of AI as a population of researchers that keeps growing with compute itself. Actually, let's go there. Can you explain the intuition that compute is a good proxy for the number of AI researchers so to speak?

Carl Shulman
So far I've talked about hardware as an initial example because we had good data about a past period. You can also make improvements on the software side and when we think about an intelligence explosion that can include — AI is doing work on making hardware better, making better software, making more hardware. But the basic idea for the hardware is especially simple in that if you have an AI worker that can substitute for a human, if you have twice as many computers you can run two separate instances of them and then they can do two different jobs, manage two different machines, work on two different design problems. Now you can get more gains than just what you would get by having two instances. We get improvements from using some of our compute not just to run more instances of the existing AI, but to train larger AIs. There's hardware technology, how much you can get per dollar you spend on hardware and there's software technology and the software can be copied freely. So if you've got the software it doesn't necessarily make that

much sense to say that — “Oh, we've got you a hundred Microsoft Windows.” You can make as many copies as you need for whatever Microsoft will charge you. But for hardware, it’s different. It matters how much we actually spend on the hardware at a given price. And if we look at the changes that have been driving AI recently, that is the thing that is really off-trend. We are spending tremendously more money on computer hardware for training big AI models.

Dwarkesh Patel
Okay so there's the investment in hardware, there's the hardware technology itself, and there's the software progress itself. The AI is getting better because we're spending more money on it because our hardware itself is getting better over time and because we're developing better models or better adjustments to those models. Where is the loop here?

Carl Shulman
The work involved in designing new hardware and software is being done by people now. They use computer tools to assist them, but computer time is not the primary cost for NVIDIA designing chips, for TSMC producing them, or for ASML making lithography equipment to serve the TSMC fabs. And even in AI software research that has become quite compute intensive we're still in the range where at a place like DeepMind salaries were still larger than compute for the experiments. Although more recently tremendously more of the expenditures were on compute relative to salaries. If you take all the work that's being done by those humans, there's like low tens of thousands of people working at Nvidia designing GPUs specialized for AI. There's more than 70,000 people at TSMC which is the leading producer of cutting-edge chips. There's a lot of additional people at companies like ASML that supply them with the tools they need and then a company like DeepMind, I think from their public filings, they recently had a thousand people. OpenAI is a few hundred people. Anthropic is less. If you add up things like Facebook AI research, Google Brain, other R&D, you get thousands or tens of thousands of people who are working on AI research.

We would want to zoom in on those who are developing new methods rather than narrow applications. So inventing the transformer definitely counts but optimizing for some particular businesses data set cleaning probably not. So those people are doing this work, they're driving quite a lot of progress. What we observe in the growth of people relative to the growth of those capabilities is that pretty consistently the capabilities are doubling on a shorter time scale than the people required to do them are doubling. We talked about hardware and how it was pretty dramatic historically. Like four or five doublings of compute efficiency per doubling of human inputs. I think that's a bit lower now as we get towards the end of Moore's law although interestingly not as much lower as you might think because the growth of inputs has also slowed recently. On the software side there's some work by Tamay Besiroglu and collaborators; it may have been his thesis. It's called Are models getting harder to find? and it's applying the same analysis as the “Are ideas getting harder to find?” and you can look at growth rates of papers, from citations, employment at these companies, and it seems like the doubling time of these like workers driving the software advances is like several years whereas the doubling of effective compute from algorithmic progress is faster. There's a group called Epoch, they've received grants from open philanthropy, and they do work collecting datasets that are relevant to forecasting AI progress. Their headline results for what's the rate of progress in hardware and software, and growth in budgets are as follows — For hardware, they're looking at a doubling of hardware efficiency in like two years. It's possible it's a bit better than that when you take into account certain specializations for AI workloads. For the growth of budgets they find a doubling time that's something like six months in recent years which is pretty tremendous relative to the historical rates. We should maybe get into that later and then on the algorithmic progress side, mainly using Imagenet type datasets right now they find a doubling time that's less than one year. So when you combine all of these things the growth of effective compute for training big AIs is pretty drastic.

Dwarkesh Patel
I think I saw an estimate that GPT-4 cost like 50 million dollars or around that range to train. Now suppose that AGI takes a 1000x that, if you were just a scale of GPT-4 it might not be that but just for the sake of example, some part of that will come from companies just spending a lot more to train the models and that's just greater investment. Part of that will come from them having better models.You get the same effect of increasing it by 10x just from having a better model. You can spend more money on it to train a bigger model, you can just have a better model, or you can have chips that are cheaper to train so you get more compute for the same dollars. So those are the three you are describing the ways in which the "effective compute" would increase?

Carl Shulman
Looking at it right now, it looks like you might get two or three doublings of effective compute for this thing that we're calling software progress which people get by asking — how much less compute can you use now to achieve the same benchmark as you achieved before? There are reasons to not fully identify this with software progress as you might naively think because some of it can be enabled by the other. When you have a lot of compute you can do more experiments and find algorithms that work better. We were talking earlier about how sometimes with the additional compute you can get higher efficiency by running a bigger model. So that means you're getting more for each GPU that you have because you made this larger expenditure. That can look like a software improvement because this model is not a hardware improvement directly because it's doing more with the same hardware but you wouldn't have been able to achieve it without having a ton of GPUs to do the big training run.

Dwarkesh Patel
The feedback loop itself involves the AI that is the result of this greater effect of compute helping you train better AI or use less effective compute in the future to train better AI?

Carl Shulman

It can help with the hardware design. NVIDIA is a fab-less chip design company. They don't make their own chips. They send files of instructions to TSMC which then fabricates the chips in their own facilities. If you could automate the work of those 10,000+ people and have the equivalent of a million people doing that work then you would pretty quickly get the kind of improvements that can be achieved with the existing nodes that TSMC is operating on and get a lot of those chip design gains. Basically doing the job of improving chip design that those people are working on now but get it done faster. While that's one thing I think that's less important for the intelligence explosion. The reason being that when you make an improvement to chip design it only applies to the chips you make after that. If you make an improvement in AI software, it has the potential to be immediately applied to all of the GPUs that you already have. So the thing that I think is most disruptive and most important and has the leading edge of the change from AI automation of the inputs to AI is on the software side.

Dwarkesh Patel

At what point would it get to the point where the AIs are helping develop better software or better models for future AIs? Some people claim today, for example, that programmers at OpenAI are using Copilot to write programs now. So in some sense you're already having that feedback loop but I'm a little skeptical of that as a mechanism. At what point would it be the case that the AI is contributing significantly in the sense that it would almost be the equivalent of having additional researchers to AI progress and software?

Carl Shulman

The quantitative magnitude of the help is absolutely central. There are plenty of companies that make some product that very slightly boosts productivity. When Xerox makes fax machines, it maybe increases people's productivity in office work by 0.1% or something. You're not gonna have explosive growth out of that because 0.1% more effective R&D at Xerox and any customers buying the machines is not that important. The thing to look for is — when is it the case that the contributions from AI are starting to become as large as the contributions from humans? So when this is boosting their effective productivity by 50 or 100% and if you then go from like eight months doubling time for effective compute from software innovations, things like inventing the transformer or discovering chinchilla scaling and doing your training runs more optimally or creating flash attention. If you move that from 8 months to 4 months and then the next time you apply that it significantly increases the boost you're getting from the AI. Now maybe instead of giving a 50% or 100% productivity boost now it's more like 200%. It doesn't have to have been able to automate everything involved in the process of AI research. It can be that it's automated a bunch of things and then those are being done in extreme profusion. A thing AI can do, you can have it done much more often because it's so cheap. And so it's not a threshold of — this is human level AI, it can do everything a human can do with no weaknesses in any area. It's that, even with its weaknesses it's able to bump up the performance. So that instead of getting the results we would have with the 10,000 people working on finding these innovations, we get the results that we would have if we had twice as many of those people with the same kind of skill distribution.

It’s a demanding challenge, you need quite a lot of capability for that but it's also important that it's significantly less than — this is a system where there's no way you can point at it and say in any respect it is weaker than a human. A system that was just as good as a human in every respect but also had all of the advantages of an AI, that is just way beyond this point. If you consider that the output of our existing fabs make tens of millions of advanced GPUs per year. Those GPUs if they were running AI software that was as efficient as humans, it is sample efficient, it doesn't have any major weaknesses, so they can work four times as long, the 168 hour work week, they can have much more education than any human. A human, you got a PhD, it's like 20 years of education, maybe longer if they take a slow route on the PhD. It's just normal for us to train large models by eat the internet, eat all the published books ever, read everything on GitHub and get good at predicting it. So the level of education vastly beyond any human, the degree to which the models are focused on task is higher than all but like the most motivated humans when they're really, really gunning for it. So you combine the things tens of millions of GPUs, each GPU is doing the work of the very best humans in the world and the most capable humans in the world can command salaries that are a lot higher than the average and particularly in a field like STEM or narrowly AI, like there's no human in the world who has a thousand years of experience with TensorFlow or let alone the new AI technology that was invented the year before but if they were around, yeah, they'd be paid millions of dollars a year. And so when you consider this — tens of millions of GPUs. Each is doing the work of 40, maybe more of these existing workers, is like going from a workforce of tens of thousands to hundreds of millions. You immediately make all kinds of discoveries, then you immediately develop all sorts of tremendous technologies. Human level AI is deep, deep into an intelligence explosion. Intelligence explosion has to start with something weaker than that.

Dwarkesh Patel
Yeah, what is the thing it starts with and how close are we to that? Because to be a researcher at OpenAI is not just completing the hello world Prompt that Copilot does right? You have to choose a new idea, you have to figure out the right way to approach it, you perhaps have to manage the people who are also working with you on that problem. It's an incredibly complicated portfolio of skills rather than just a single skill. What is the point at which that feedback loop starts where you're not just doing the 0.5% increase in productivity that an AI tool might do but is actually the equivalent of a researcher or close to it?

Carl Shulman
Maybe a way is to give some illustrative examples of the kinds of capabilities that you might see. Because these systems have to be a lot weaker than the human-level things, what we'll have is intense application of the ways in which AIs have advantages partly offsetting their weaknesses. AIs are cheap so we can call a lot of them to do many small problems. You'll have situations where you have dumber AIs that are deployed thousands of times to equal one human worker. And they'll be doing things like voting algorithms where with an LLM you generate a bunch of different responses and take a majority vote among them that improves some performance. You'll have things like the AlphaGo kind of approach where you use the neural net to do search and you go deeper with the search by plowing in more compute which helps to offset the inefficiency and weaknesses of the model on its own. You'll do things that would just be totally impractical for humans because of the sheer number of steps, an example of that would be designing synthetic training data. Humans do not learn by just going into the library and opening books at random pages, it's actually much much more efficient to have things like schools and classes where they teach you things in an order that makes sense, focusing on the skills that are more valuable to learn. They give you tests and exams. They're designed to try and elicit the skill they're actually trying to teach. And right now we don't bother with that because we can hoover up more data from the internet. We're getting towards the end of that but yeah, as the AIs get more sophisticated they'll be better able to tell what is a useful kind of skill to practice and to generate that. We've done that in other areas like AlphaGo. The original version of AlphaGo was booted up with data from human Go play and then improved with reinforcement learning and Monte-carlo tree search but then AlphaZero, a somewhat more sophisticated model benefited from some other improvements but was able to go from scratch and it generated its own data through self play. Getting data of a higher quality than the human data because there are no human players that good available in the data set and also a curriculum so that at any given point it was playing games against an opponent of equal skill itself. It was always in an area when it was easy to learn. If you're just always losing no matter what you do, or always winning no matter what you do, it's hard to distinguish which things are better and which are worse? And when we have somewhat more sophisticated AIs that can generate training data and tasks for themselves, for example if the AI can generate a lot of unit tests and then can try and produce programs that pass those unit tests, then the interpreter is providing a training signal and the AI can get good at figuring out what's the kind of programming problem that is hard for AIs right now that will develop more of the skills that I need and then do them. You're not going to have employees at Open AI write a billion programming problems, that's just not gonna happen. But you are going to have AIs given the task of producing the enormous number of programming challenges.

Dwarkesh Patel
In LLMs themselves, there's a paper out of Anthropic called Constitution AI where they basically had the program just talk to itself and say, "Is this response helpful? If not, how can I make this more helpful" and the responses improved and then you train the model on the

more helpful responses that it generates by talking to itself so that it generates it natively and you could imagine more sophisticated or better ways to do that. But then the question is GPT-4 already costs like 50 million or 100 million or whatever it was. Even if we have greater effective compute from hardware increases and better models, it's hard to imagine how we could sustain four or five orders of magnitude greater effective size than GPT-4 unless we're dumping in trillions of dollars, the entire economies of big countries, into training the next version. The question is do we get something that can significantly help with AI progress before we run out of the sheer money and scale and compute that would require to train it? Do you have a take on that?

Carl Shulman
First I'd say remember that there are these three contributing trends. The new H100s are significantly better than the A100s and a lot of companies are actually just waiting for their deliveries of H100s to do even bigger training runs along with the work of hooking them up into clusters and engineering the thing. All of those factors are contributing and of course mathematically yeah, if you do four orders of magnitude more than 50 or 100 million then you're getting to trillion dollar territory. I think the way to look at it is at each step along the way, does it look like it makes sense to do the next step? From where we are right now seeing the results with GPT-4 and ChatGPT companies like Google and Microsoft are pretty convinced that this is very valuable. You have talk at Google and Microsoft that it's a billion dollar matter to change market share in search by a percentage point so that can fund a lot. On the far end if you automate human labor we have a hundred trillion dollar economy and most of that economy is paid out in wages, between 50 and 70 trillion dollars per year. If you create AGI it's going to automate all of that and keep increasing beyond that. So the value of the completed project Is very much worth throwing our whole economy into it, if you're going to get the good version and not the catastrophic destruction of the human race or some other disastrous outcome. In between it's a question of — how risky and uncertain is the next step and how much is the growth in revenue you can generate with it? For moving up to a billion dollars I think that's absolutely going to happen. These large tech companies have R&D budgets of tens of billions of dollars and when you think about it in the relevant sense all the employees at Microsoft who are doing software engineering that's contributing to creating software objects, it's not weird to spend tens of billions of dollars on a product that would do so much. And I think that it's becoming clearer that there is a market opportunity to fund the thing. Going up to a hundred billion dollars, that's the existing R&D budgets spread over multiple years. But if you keep seeing that when you scale up the model it substantially improves the performance, it opens up new applications, that is you're not just improving your search but maybe it makes self-driving cars work, you replace bulk software engineering jobs or if not replace them amplify productivity. In this kind of dynamic you actually probably want to employ all the software engineers you can get as long as they are able to make any contribution because the returns of improving stuff in AI itself gets so high. But yeah, I think that can go up to a hundred billion. And at a hundred billion you're using a significant fraction of our existing fab capacity. Right now the revenue

of NVIDIA is 25 billion, the revenue of TSMC is over 50 billion. I checked in 2021, NVIDIA was maybe 7.5%, less than 10% of TSMC revenue. So there's a lot of room and most of that was not AI chips. They have a large gaming segment, there are data center GPU's that are used for video and the like. There's room for more than an order of magnitude increase by redirecting existing fabs to produce more AI chips and they're just actually using the AI chips that these companies have in their cloud for the big training runs. I think that that's enough to go to the 10 billion and then combine with stuff like the H100 to go up to the hundred billion.

Dwarkesh Patel
Just to emphasize for the audience the initial point about revenue made. If it costs OpenAI 100 million dollars to train GPT-4 and it generates 500 million dollars in revenue, you pay back your expenses with 100 million and you have 400 million for your next training run. Then you train your GPT 4.5, you get let's say four billion dollars in revenue out of that. That's where the feedback group of revenue comes from. Where you're automating tasks and therefore you're making money you can use that money to automate more tasks. On the ability to redirect the fab production towards AI chips, fabs take a decade or so to build. Given the ones we have now and the ones that are going to come online in the next decade, is there enough to sustain a hundred billion dollars of GPU compute if you wanted to spend that on a training run?

Carl Shulman
Yes, you definitely make the hundred billion one. As you go up to a trillion dollar run and larger, it's going to involve more fab construction and yeah, fabs can take a long a long time to build. On the other hand, if in fact you're getting very high revenue from the AI systems and you're actually bottlenecked on the construction of these fabs then their price could skyrocket and that could lead to measures we've never seen before to expand and accelerate fab production. If you consider, at the limit you're getting models that approach human-like capability, imagine things that are getting close to brain-like efficiencies plus AI advantages. We were talking before a cluster of GPU supporting AIs that do things, data parallelism. If that can work four times as much as a highly skilled motivated focused human with levels of education that have never been seen in the human population, and if a typical software engineer can earn hundreds of thousands of dollars, the world's best software engineers can earn millions of dollars today and maybe more in a world where there's so much demand for AI. And then times four for working all the time. If you can generate close to 10 million dollars a year out of the future version H100 and it cost tens of thousands of dollars with a huge profit margin now. And profit margin could be reduced with large production. That is a big difference that that chip pays for itself almost instantly and you could support paying 10 times as much to have these fabs constructed more rapidly. If AI is starting to be able to contribute more of the skilled technical work that makes it hard for NVIDIA to suddenly find thousands upon thousands of top quality engineering hires.

If AI hasn't reached that level of performance then this is how you can have things stall out. A world where AI progress stalls out is one where you go to the 100 billion and then over succeeding years software progress turns out to stall. You lose the gains that you are getting from moving researchers from other fields. Lots of physicists and people from other areas of computer science have been going to AI but you tap out those resources as AI becomes a larger proportion of the research field. And okay, you've put in all of these inputs, but they just haven't yielded AGI yet. I think that set of inputs probably would yield the kind of AI capabilities needed for intelligence explosion but if it doesn't, after we've exhausted this current scale up of increasing the share of our economy that is trying to make AI. If that's not enough then after that you have to wait for the slow grind of things like general economic growth, population growth and such and so things slow. That results in my credences and this kind of advanced AI happening to be relatively concentrated, over the next 10 years compared to the rest of the century because we can't keep going with this rapid redirection of resources into AI. That's a one-time thing.

Dwarkesh Patel
If the current scale up works we're going to get to AGI really fast, like within the next 10 years or something. If the current scale up doesn't work, all we're left with is just like the economy growing 2% a year, we have 2% a year more resources to spend on AI and at that scale you're talking about decades before just through sheer brute force you can train the 10 trillion dollar model or something. Let's talk about why you have your thesis that the current scale up would work. What is the evidence from AI itself or maybe from primate evolution and the evolution of other animals? Just give me the whole confluence of reasons that make you think that.

Carl Shulman
Maybe the best way to look at that might be to consider, when I first became interested in this area, so in the 2000s which was before the deep learning revolution, how would I think about timelines? How did I think about timelines? And then how have I updated based on what has been happening with deep learning? Back then I would have said we know the brain is a physical object, an information processing device, it works, it's possible and not only is it possible it was created by evolution on earth. That gives us something of an upper bound in that this kind of brute force was sufficient. There are some complexities like what if it was a freak accident and that didn't happen on all of the other planets and that added some value. I have a paper with Nick Bostrom on this. I think basically that's not that important an issue. There's convergent evolution, octopi are also quite sophisticated. If a special event was at the level of forming cells at all, or forming brains at all, we get to skip that because we're choosing to build computers and we already exist. We have that advantage. So evolution gives something of an upper bound, really intensive massive brute force search and things like evolutionary algorithms can produce intelligence.

Dwarkesh Patel
Isn’t the fact that octopi and other mammals got to the point of being pretty intelligent but not human level intelligent some evidence that there's a hard step between a cephalopod and a human?

Carl Shulman
Yeah, that would be a place to look but it doesn't seem particularly compelling. One source of evidence on that is work by Herculano-Houzel. She's a neuroscientist who has dissolved the brains of many creatures and by counting the nuclei she's able to determine how many neurons are present in different species and has found a lot of interesting trends in scaling laws. She has a paper discussing the human brain as a scaled up primate brain. Across a wide variety of animals, mammals in particular, there's certain characteristic changes in the number of neurons and the size of different brain regions as things scale up. There's a lot of structural similarity there and you can explain a lot of what is different about us with a brute force story which is that you expend resources on having a bigger brain, keeping it in good order, and giving it time to learn. We have an unusually long childhood. We spend more compute by having a larger brain than other animals, more than three times as large as chimpanzees, and then we have a longer childhood than chimpanzees and much more than many, many other creatures. So we're spending more compute in a way that's analogous to having a bigger model and having more training time with it. And given that we see with our AI models, these large consistent benefits from increasing compute spent in those ways and with qualitatively new capabilities showing up over and over again particularly in areas that AI skeptics call out. In my experience over the last 15 years the things that people call out are like —“Ah, but the AI can't do that and it's because of a fundamental limitation.” We've gone through a lot of them. There were Winograd schemas, catastrophic forgetting, quite a number and they have repeatedly gone away through scaling. So there's a picture that we're seeing supported from biology and from our experience with AI where you can explain — Yeah, in general, there are trade-offs where the extra fitness you get from a brain is not worth it and so creatures wind up mostly with small brains because they can save that biological energy and that time to reproduce, for digestion and so on. Humans seem to have wound up in a self-reinforcing niche where we greatly increase the returns to having large brains. Language and technology are the obvious candidates. You have humans around you who know a lot of things and they can teach you. And compared to almost any other species we have vastly more instruction from parents and the society of the [unclear]. You're getting way more from your brain than you get per minute because you can learn a lot more useful skills and then you can provide the energy you need to feed that brain by hunting and gathering, by having fire that makes digestion easier.

Basically how this process goes on is that it's increasing the marginal increase in reproductive fitness you get from allocating more resources along a bunch of dimensions towards cognitive ability. That's bigger brains, longer childhood, having our attention be more on learning. Humans play a lot and we keep playing as adults which is a very weird thing compared to other animals. We're more motivated to copy other humans around us than the other primates. These are motivational changes that keep us using more of our attention and effort on learning which pays off more when you have a bigger brain and a longer lifespan in which to learn in.

Many creatures are subject to lots of predation or disease. If you're mayfly or a mouse and if you try and invest in a giant brain and a very long childhood you're quite likely to be killed by some predator or some disease before you're actually able to use it. That means you actually have exponentially increasing costs in a given niche. If I have a 50% chance of dying every few months, as a little mammal or a little lizard, that means the cost of going from three months to 30 months of learning and childhood development is not 10 times the loss, it’s 2^-10. A factor of 1024 reduction in the benefit I get from what I ultimately learn because 99.9 percent of the animals will have been killed before that point. We're in a niche where we're a large long-lived animal with language and technology so where we can learn a lot from our groups. And that means it pays off to just expand our investment on these multiple fronts in intelligence.

Dwarkesh Patel
That's so interesting. Just for the audience the calculation about like two to the whatever months is just like, you have a half chance of dying this month, a half chance of dying next month, you multiply those together. There's other species though that do live in flocks or as packs. They do have a smaller version of the development of cubs that play with each other. Why isn't this a hill on which they could have climbed to human level intelligence themselves? If it's something like language or technology, humans were getting smarter before we got language. It seems like there should be other species that should have beginnings of this cognitive revolution especially given how valuable it is given we've dominated the world. You would think there would be selective pressure for it.

Carl Shulman
Evolution doesn't have foresight. The thing in this generation that gets more surviving offspring and grandchildren is the thing that becomes more common. Evolution doesn't look ahead and think oh in a million years you'll have a lot of descendants. It's what survives and reproduces now. In fact, there are correlations where social animals do on average have larger brains and part of that is probably the additional social applications of brains, like keeping track of which of your group members have helped you before so that you can reciprocate. You scratch my back, I'll scratch yours. Remembering who's dangerous within the group is an additional application of intelligence. So there's some correlation there but what it seems like is that in most of these cases it's enough to invest more but not invest to the point where a mind can easily develop language and technology and pass it on. You see bits of tool use in some other primates who have an advantage compared to say whales who have quite large brains partly because they are so large themselves and they have some other things, but they don't have hands which means that reduces a bunch of ways in which

brains can pay off and investments in the functioning of that brain. But yeah, primates will use sticks to extract termites, Capuchin monkeys will open clams by smashing them with a rock. But what they don't have is the ability to sustain culture. A particular primate will maybe discover one of these tactics and it'll be copied by their immediate group but they're not holding on to it that well. When they see the other animal do it they can copy it in that situation but they don't actively teach each other in their population. So it's easy to forget things, easy to lose information and in fact they remain technologically stagnant for hundreds of thousands of years.

And we can look at some human situations. There's an old paper, I believe by the economist Michael Kramer, which talks about technological growth in the different continents for human societies. Eurasia is the largest integrated connected area. Africa is partly connected to it but the Sahara desert restricts the flow of information and technology and such. Then you have the Americas after the colonization from the land bridge were largely separated and are smaller than Eurasia, then Australia, and then you had smaller island situations like Tasmania. Technological progress seems to have been faster the larger the connected group of people. And in the smallest groups, like Tasmania where you had a relatively small population, they actually lost technology. They lost some fishing techniques. And if you have a small population and you have some limited number of people who know a skill and they happen to die or there's some change in circumstances that causes people not to practice or pass on that thing then you lose it. If you have few people you're doing less innovation and the rate at which you lose technologies to some local disturbance and the rate at which you create new technologies can wind up imbalanced. The great change of hominids and humanity is that we wound up in this situation where we were accumulating faster than we were losing and accumulating those technologies allowed us to expand our population. They created additional demand for intelligence so our brains became three times as large as chimpanzees and our ancestors who had a similar brain size.

Dwarkesh Patel
Okay. And the crucial point in relevance to AI is that the selective pressures against intelligence in other animals are not acting against these neural networks because they're not going to get eaten by a predator if they spend too much time becoming more intelligent, we're explicitly training them to become more intelligent. So we have good first principles reason to think that if it was scaling that made our minds this powerful and if the things that prevented other animals from scaling are not impinging on these neural networks, these things should just continue to become very smart.

Carl Shulman
Yeah, we are growing them in a technological culture where there are jobs like software engineer that depend much more on cognitive output and less on things like metabolic resources devoted to the immune system or to building big muscles to throw spears.

Dwarkesh Patel
This is kind of a side note but I'm just kind of interested. You referenced Chinchilla scaling at some point. For the audience this is a paper from DeepMind which describes if you have a model of a certain size what is the optimum amount of data that it should be trained on? So you can imagine bigger models, you can use more data to train them and in this way you can figure out where you should spend your compute. Should you spend it on making the model bigger or should you spend it on training it for longer? In the case of different animals, in some sense how big their brain is like model sizes and they're training data sizes like how long they're cubs or how long their infants or toddlers before they're full adults. I'm curious, is there some kind of scaling law?

Carl Shulman
Chinchilla scaling is interesting because we were talking earlier about the cost function for having a longer childhood where it's exponentially increasing in the amount of training compute you have when you have exogenous forces that can kill you. Whereas when we do big training runs, the cost of throwing in more GPU is almost linear and it's much better to be linear than exponentially decay as you expend resources.

Dwarkesh Patel
Oh, that's a really good point.

Carl Shulman
Chinchilla scaling would suggest that for a brain of human size it would be optimal to have many millions of years of education but obviously that's impractical because of exogenous mortality for humans. So there's a fairly compelling argument that relative to the situation where we would train AI that animals are systematically way under trained. They're more efficient than our models. We still have room to improve our algorithms to catch up with the efficiency of brains but they are laboring under that disadvantage.

Dwarkesh Patel
That is so interesting. I guess another question you could have is: Humans got started on this evolutionary hill climbing route where we're getting more intelligent because it has more benefits for us. Why didn't we go all the way on that route? If intelligence is so powerful why aren't all humans as smart as we know humans can be? If intelligence is so powerful, why hasn't there been stronger selective pressure? I understand hip size, you can't give birth to a really big headed baby or whatever. But you would think evolution would figure out some way to offset that if intelligence has such big power and is so useful.

Carl Shulman
Yeah, if you actually look at it quantitatively that's not true and even in recent history it looks like a pretty close balance between the costs and the benefits of having more cognitive abilities. You say, who needs to worry about the metabolic costs? Humans put 20 percent of

our metabolic energy into the brain and it's higher for young children. And then there's like breathing and digestion and the immune system. For most of history people have been dying left and right. A very large proportion of people will die of infectious disease and if you put more resources into your immune system you survive. It's life or death pretty directly via that mechanism. People die more of disease during famine and so there's boom or bust. If you have 20% less metabolic requirements [unclear] you're much more likely to survive that famine. So these are pretty big.

And then there's a trade-off about just cleaning mutational load. So every generation new mutations and errors happen in the process of reproduction. We know there are many genetic abnormalities that occur through new mutations each generation and in fact Down syndrome is the chromosomal abnormality that you can survive. All the others just kill the embryo so we never see them. But down syndrome occurs a lot and there are many other lethal mutations and there are enormous numbers of less damaging mutations that are degrading every system in the body. Evolution each generation has to pull away at some of this mutational load and the priority with which that mutational load is pulled out scales in proportion to how much the traits it is affecting impact fitness. So you get new mutations that impact your resistance to malaria, you got new mutations that damage brain function and then those mutations are purged each generation. If malaria is a bigger difference in mortality than the incremental effectiveness as a hunter-gatherer you get from being slightly more intelligent, then you'll purge that mutational load first. Similarly humans have been vigorously adapting to new circumstances. Since agriculture people have been developing things like the ability to have amylase to digest breads and milk. If you're evolving for all of these things and if some of the things that give an advantage for that incidentally carry along nearby them some negative effect on another trait then that other trait can be damaged. So it really matters how important to survival and reproduction cognitive abilities were compared to everything else the organism has to do. In particular, surviving famine, having the physical abilities to do hunting and gathering and even if you're very good at planning your hunting, being able to throw a spear harder can be a big difference and that needs energy to build those muscles and then to sustain them.

Given all these factors it's not a slam dunk to invest at the margin. And today, having bigger brains is associated with greater cognitive ability but it's modest. Large-scale pre-registered studies with MRI data. The correlation is in a range of 0.25 - 0.3 and the standard deviation of brain size is like 10%. So if you double the size of the brain, the existing brain costs like 20 of metabolic energy go up to 40%, okay, that's like eight standard deviations of brain size if the correlation is 0.25 then yeah, you get a gain from that eight standard deviations of brain size, two standard deviations of cognitive ability. In our modern society, where cognitive ability is very rewarded and finishing school and becoming an engineer or a doctor or whatever can pay off a lot financially, the average observed return in income is still only one or two percent proportional increase. There's more effects at the tail, there's more effect in professions like STEM but on the whole it's not a lot. If it was like a five percent increase or a 10 percent increase then you could tell a story where yeah, this is hugely increasing the amount of food you could have, you could support more children, but it's a modest effect and the metabolic costs will be large and then throw in these other these other aspects. Else we can just see there was not very strong rapid directional selection on the thing which would be there if by solving a math puzzle you could defeat malaria, then there would be more evolutionary pressure.

Dwarkesh Patel
That is so interesting. Not to mention of course that if you had 2x the brain size, without c-section you or your mother or both would die. This is a question I've actually been curious about for over a year and I've briefly tried to look up an answer. I know this was off topic and my apologies to the audience, but I was super interested and that was the most comprehensive and interesting answer I could have hoped for. So yeah, we have a good explanation or good first principles evolution or reason for thinking that intelligence scaling up to humans is not implausible just by throwing more scale at it.

Carl Shulman
I would also add that we also have the brain right here with us available for neuroscience to reverse engineer its properties. This was something that would have mattered to me more in the 2000s. Back then when I said, yeah, I expect this by the middle of the century-ish, that was a backstop if we found it absurdly difficult to get to the algorithms and then we would learn from neuroscience. But in actual history, it's really not like that. We develop things in AI and then also we can say oh, yeah, this is like this thing in neuroscience or maybe this is a good explanation. It's not as though neuroscience Is driving AI progress. It turns out not to be that necessary.

Dwarkesh Patel
I guess that is similar to how planes were inspired by the existence proof of birds but jet engines don't flap. All right, good reason to think scaling might work. So we spent a hundred billion dollars and we have something that is like human level or can help significantly with AI research.

Carl Shulman
I mean that that might be on the earlier end but I definitely would not rule that out given the rates of change we've seen with the last few scale ups.

Dwarkesh Patel
At this point somebody might be skeptical. We already have a bunch of human researchers, how profitable is the incremental researcher? And then you might say no, this is thousands of researchers. I don't know how to express this skepticism exactly. But skeptical of just generally the effect of scaling up the number of people working on the problem to rapid-rapid progress on that problem. Somebody might think that with humans the reason

the amount of population working on a problem is such a good proxy for progress on the problem is that there's already so much variation that is accounted for. When you say there's a million people working on a problem, there's hundreds of super geniuses working on it, thousands of people who are very smart working on it. Whereas with an AI all the copies are the same level of intelligence and if it's not super genius intelligence the total quantity might not matter as much.

Carl Shulman
I'm not sure what your model is here. Is the model that the diminishing returns kickoff, suddenly has a cliff right where we are? There were results in the past from throwing more people at problems and this has been useful in historical prediction, this idea of experience curves and [unclear] law measuring cumulative production in a field, which is also going to be a measure of the scale of effort and investment, and people have used this correctly to argue that renewable energy technology, like solar, would be falling rapidly in price because it was going from a low base of very small production runs, not much investment in doing it efficiently, and climate advocates correctly called out, people like David Roberts, the futurist [unclear] actually has some interesting writing on this. They correctly called out that there would be a really drastic fall in prices of solar and batteries because of the increasing investment going into that. The human genome project would be another. So I’d say there's real evidence. These observed correlations, from ideas getting harder to find, have held over a fair range of data and over quite a lot of time. So I'm wondering what’s the nature of the deviation you're thinking of?

Dwarkesh Patel
Maybe this is a good way to describe what happens when more humans enter a field but does it even make sense to say that a greater population of AIs is doing AI research if there's like more GPUs running a copy of GPT-6 doing AI research. How applicable are these economic models of the quantity of humans working on a problem to the magnitude of AIs working on a problem?

Carl Shulman
If you have AIs that are directly automating particular jobs that humans were doing before then we say, well with additional compute we can run more copies of them to do more of those tasks simultaneously. We can also run them at greater speed. Some people have an intuition that what matters is time, that it's not how many people working on a problem at a given point. I think that doesn't bear out super well but AI can also run faster than humans. If you have a set of AIs that can do the work of the individual human researchers and run at 10 times or 100 times the speed. And we ask well, could the human research community have solved these algorithm problems, do things like invent transformers over 100 years, if we have AIs with a population effective population similar to the humans but running 100 times as fast and so. You have to tell a story where no, the AI can't really do the same things as the

humans and we're talking about what happens when the AIs are more capable of in fact doing that.

Dwarkesh Patel
Although they become more capable as lesser capable versions of themselves help us make themselves more capable, right? You have to kickstart that at some point. Is there an example in analogous situations? Is intelligence unique in the sense that you have a feedback loop of — with a learning curve or something else, a system’s outputs are feeding into its own inputs. Because if we're talking about something like Moore's law or the cost of solar, you do have this way where we're throwing more people with the problem and we're making a lot of progress, but we don't have this additional part of the model where Moore's law leads to more humans somehow and more humans are becoming researchers.

Carl Shulman
You do actually have a version of that in the case of solar. You have a small infant industry that's doing things like providing solar panels for space satellites and then getting increasing amounts of subsidized government demand because of worries about fossil fuel depletion and then climate change. You can have the dynamic where visible successes with solar and lowering prices then open up new markets. There's a particularly huge transition where renewables become cheap enough to replace large chunks of the electric grid. Earlier you were dealing with very niche situations like satellites, it’s very difficult to refuel a satellite in place and in remote areas. And then moving to the sunniest areas in the world with the biggest solar subsidies. There was an element of that where more and more investment has been thrown into the field and the market has rapidly expanded as the technology improved. But I think the closest analogy is actually the long run growth of human civilization itself and I know you had Holden Karnofsky from the open philanthropy project on earlier and discuss some of this research about the long run acceleration of human population and economic growth. Developing new technologies allowed the human population to expand and humans to occupy new habitats and new areas and then to invent agriculture to support the larger populations and then even more advanced agriculture in the modern industrial society. So there, the total technology and output allowed you to support more humans who then would discover more technology and continue the process. Now that was boosted because on top of expanding the population the share of human activity that was going into invention and innovation went up and that was a key part of the industrial revolution. There was no such thing as a corporate research lab or an engineering university prior to that. So you're both increasing the total human population and the share of it going in. But this population dynamic is pretty analogous. Humans invent farming, they can have more humans, they can invent industry and so on.

Dwarkesh Patel
Maybe somebody would be skeptical that with AI progress specifically, it’s not just a matter of some farmer figuring out crop rotation or some blacksmith figuring out how to do

metallurgy better. In fact even to make the 50% improvement in productivity you basically need something on the IQ that's close to Ilya Sutskever. There's like a discontinuous line. You're contributing very little to productivity and then you're like Ilya and then you contribute a lot. You see what I'm saying? There isn't a gradual increase in capabilities that leads to the feedback.

Carl Shulman
You're imagining a case where the distribution of tasks is such that there's nothing that individually automating it particularly helps and so the ability to contribute to AI research is really end loaded. Is that what you're saying?

Dwarkesh Patel
Yeah, we already see this in these really high IQ companies or projects. Theoretically I guess Jane Street or OpenAI could hire like a bunch of mediocre people with a comparative advantage to do some menial task and that could free up the time of the really smart people but they don't do that right? Due to transaction costs or whatever else.

Carl Shulman
Self-driven cars would be another example where you have a very high quality threshold. Your performance as a driver is worse than a human, like you have 10 times the accident rate or 100 times the accident rate, then the cost of insurance for that which is a proxy for people's willingness to ride the car would be such that the insurance costs would absolutely dominate. So even if you have zero labor cost, it is offset by the increased insurance costs. There are lots of cases like that where partial automation is in practice not very usable because complementing other resources you're gonna use those other resources less efficiently. In a post-AGI future the same thing can apply to humans. People can say, comparative advantage, even if AIs can do everything better than a human well it's still worth something. Human can do something. They can lift a box, that's something. [unclear] In such an economy you wouldn't want to let a human worker into any industrial environment because in a clean room they'll be emitting all kinds of skin cells and messing things up. You need to have an atmosphere there. You need a bunch of supporting tools and resources and materials and those supporting resources and materials will do a lot more productively working with AI and robots rather than a human. You don't want to let a human anywhere near the thing just like you wouldn't want a Gorilla wandering around in a China shop. Even if you've trained it to, most of the time pick up a box for you if you give it a banana. It's just not worth it to have it wandering around your china shop.

Dwarkesh Patel
Yeah. Why is that not a good objection?

Carl Shulman
I think that that is one of the ways in which partial automation can fail to really translate into a lot of economic value. That's something that will attenuate as we go on and as the AI is more able to work independently and more able to handle its own screw-ups and get more reliable.

Dwarkesh Patel
But the way in which it becomes more reliable is by AI progress speeding up which happens if AI can contribute to it but if there is some reliability bottleneck that prevents it from contributing to that progress then you don't have the loop, right?

Carl Shulman
I mean this is why we're not there yet.

Dwarkesh Patel
But then what is the reason to think we'll be there?

Carl Shulman
The broad reason is the inputs are scaling up. Epoch have a paper called compute trends across three eras of machine learning and they look at the compute expended on machine learning systems since the founding of the field of AI, the beginning of the 1950s. Mostly it grows with Moore's law and so people are spending a similar amount on their experiments but they can just buy more with that because the compute is coming. That data covers over 20 orders of magnitude, maybe like 24, and of all of those increases since 1952 a little more than half of them happened between 1952 and 2010 and all the rest since 2010. We've been scaling that up four times as fast as was the case for most of the history of AI. We're running through the orders of magnitude of possible resource inputs you could need for AI much much more quickly than we were for most of the history of AI. That's why this is a period with a very elevated chance of AI per year because we're moving through so much of the space of inputs per year and indeed it looks like this scale-up taken to its conclusion will cover another bunch of orders of magnitude and that's actually a large fraction of those that are left before you start running into saying well, this is going to have to be like evolution with the simple hacks we get to apply. We're selecting for intelligence the whole time, we're not going to do the same mutation that causes fatal childhood cancer a billion times even though I mean we keep getting the same fatal mutations even though they've been done many times. We use gradient descent which takes into account the derivative of improvement on the loss all throughout the network and we don't throw away all the contents of the network with each generation where you compress down to a little DNA. So there's that bar of, if you're going to do brute force like evolution combined with these very simple ways we can save orders of magnitude on that. We're going to cover a fraction that's like half of that distance in this scale-up over the next 10 years or so. And so if you started off with a kind of vague uniform prior, you probably can't make AGI with the amount of

compute that would be involved in a fruit fly existing for a minute which would be the early days of AI. Maybe you would get lucky, we were able to make calculators because calculators benefited from very reliable serially fast computers and where we could take a tiny tiny tiny tiny fraction of a human brain's compute and use it for a calculator. We couldn't take an ant's brain and rewire it to calculate. It's hard to manage ant farms let alone get them to do arithmetic for you. So there were some things where we could exploit the differences between biological brains and computers to do stuff super efficiently on computers. We would doubt that we would be able to do so much better than biology that with a tiny fraction of an insect's brain we'd be able to get AI early on. On the far end, it seemed very implausible that we couldn't do better than completely brute force evolution. And so in between you have some number of orders of magnitude of inputs where it might be. In the 2000s, I would say well, I'm gonna have a pretty uniformish prior I'm gonna put weight on it happening at the equivalent of 10^25 ops, 10^30, 10^35 and spreading out over that and then I can update another information. And in the short term, in 2005 I would say, I don't see anything that looks like the cusp of AGI so I'm also gonna lower my credence for the next five years or the next 10 years. And so that would be kind of like a vague prior and then when we take into account how quickly are we running through those orders of magnitude. If I have a uniform prior I assign half of my weight to the first half of remaining orders of magnitude and if we're gonna run through those, over the next 10 years and some, then that calls on me to put half of my credence, conditional on if ever we're gonna make AI which seems likely considering it's a material object easier than evolution, I've got to put similarly a lot of my credence on AI happening in this scale up and then that's supported by what we're seeing In terms of the rapid advances and capabilities with AI and LLMs in particular.

Dwarkesh Patel
Okay, that's actually a really interesting point. Now somebody might say, there's not some sense in which AIs could universally speed up the progress of OpenAI by 50 percent or 100 percent or 200 percent if they're not able to do everything better than Ilya Sutskever can. There's going to be something in which we're bottlenecked by the human researchers and bottleneck effects dictate that the slowest moving part of the organization will be the one that kind of determines the speed of the progress of the whole organization or the whole project. Which means that unless you get to the point where you're doing everything and everybody in the organization can do, you're not going to significantly speed up the progress of the project as a whole.

Carl Shulman
Yeah, so that is a hypothesis and I think there's a lot of truth to it. When we think about the ways in which AI can contribute, there are things we talked about before like the AI setting up their own curriculum and that's something that Ilya can't and doesn't do directly. And there's a question of how much does that improve performance? There are these things where the AI helps to produce some code for tasks and it's beyond hello world at this point.

The thing that I hear from AI researchers at leading labs is that on their core job where they're like most expert it's not helping them that much but then their job often does involve coding something that's out of their usual area of expertise or they want to research a question and it helps them there. That saves some of their time and frees them to do more of the bottlenecked work. And I think the idea of, is everything being dependent on Ilya? And is Ilya so much better than the hundreds of other employees? A lot of people who are contributing, they're doing a lot of tasks and you can have quite a lot of gain from automating some areas where you then do just an absolutely enormous amount of it relative to what you would have done before. Because things like designing the custom curriculum maybe some humans put some work into that but you're not going to employ billions of humans to produce it at scale and so it winds up being a larger share of the progress than it was before. You get some benefit from these sorts of things where there's like pieces of my job that now I can hand off to the AI and lets me focus more on the things that the AI still can't do. Later on you get to the point where yeah, the AI can do your job including the most difficult parts and maybe it has to do that in a different way. Maybe it spends a ton more time thinking about each step of a problem than you and that's the late end. The stronger these bottlenecks' effects are, the more the economic returns, the scientific returns and such are end-loaded towards getting full AGI. The weaker the bottlenecks are the more interim results will be really paying off.

Dwarkesh Patel
I probably disagree with you on how much the Ilya's of organizations seem to matter. Just from the evidence alone, how many of the big breakthroughs in deep learning was that single individual responsible for, right? And how much of his time is he spending doing anything that Copilot is helping him on? I'm guessing most of it is just managing people and coming up with ideas and trying to understand systems and so on.

And if the five or ten people who are like that at OpenAI or Anthropic or whatever, are basically the way in which algorithmic progress is happening. I know Copilot is not the thing you're talking about with like just 20% automation, but something like that. How much is that contributing to the core function of the research scientist?

Carl Shulman
Yeah, [unclear] quantitatively how much we disagree about the importance of key research employees and such. I certainly think that some researchers add more than 10 times the average employee, even much more. And obviously managers can add an enormous amount of value by proportionately multiplying the output of the many people that they manage. And so that's the kind of thing that we were discussing earlier when talking about. Well if you had a full human level AI, or AI that had all of the human capabilities plus AI advantages, you'd benchmark not off of what the typical human performance is but peak human performance and beyond. So yeah, I accept all that. I do think it makes a big difference for people how much they can outsource a lot of the tasks that are less wow, less creative and

an enormous amount is learned by experimentation. ML has been quite an experimental field and there's a lot of engineering work in building large super clusters, making hardware aware optimization and encoding of these things, being able to do the parallelism in large models, and the engineers are busy and it's not just only a big thoughts kind of area. The other branch is where will the AI advantages and disadvantages be? One AI advantage is being omnidisciplinary and familiar with the newest things. I mentioned before there's no human who has a million years of tensor flow experience. To the extent that we're interested in the very cutting edge of things that have been developed quite recently then AI that can learn about them in parallel and experiment and practice with them in parallel can potentially learn much faster than a human. And the area of computer science is one that is especially suitable for AI to learn in a digital environment so it doesn't require driving a car around that might kill someone, have enormous costs. You can do unit tests, you can prove theorems, you can do all sorts of operations entirely in the confines of a computer, which is one reason why programming has been benefiting more than a lot of other areas from LLMs recently whereas robotics is lagging. And considering they are getting better at things like the GRE, math, at programming contests, and some people have forecasts and predictions outstanding about doing well on the informatics olympiad and the Math Olympiad and in the last few years when people tried to forecast the MMLU benchmark which has a lot of sophisticated, graduate student level science kind of questions, AI knocked that down a lot faster than AI researchers and students who had registered forecasts on it. If you're getting top-notch scores on graduate exams, creative problem solving, it's not obvious that that area will be a relative weakness of AI. In fact computer science is in many ways especially suitable because of getting up to speed with new areas, being able to get rapid feedback from the interpreter at scale.

Dwarkesh Patel
But do you get rapid feedback if you're doing something that's more analogous to research? Let's say you have a new model and it's like, if we put in 10 million dollars on a mini-training run on this this would be much better.

Carl Shulman
Yeah for very large models those experiments are going to be quite expensive. You're going to look more at can you build up this capability by generalization? From things like mini math problems, programming problems, working with small networks.

Dwarkesh Patel
Yeah, fair enough. Scott Aaronson was one of my professors in college and I took his quantum information class and he recently wrote a blog post where he said, I had GPT-4 take my quantum information test and it got a B. I was like, "Damn, I got a C on the final." I updated in the direction that getting a B on a test probably means it understands quantum information pretty well.

Carl Shulman
With different areas of strengths and weaknesses than the human students.

Dwarkesh Patel
Sure, sure. Would it be possible for this intelligence explosion to happen without any hardware progress? If hardware progress stopped would this feedback loop still be able to produce some explosion with only software?

Carl Shulman
If we say that the technology is frozen, which I think is not the case right now, Nvidia has managed to deliver significantly better chips for AI workloads for the last few generations. H100, A100, V100. If that stops entirely, maybe we'll define this as no more nodes, Moore’s law is over, at that point the gains you get an amount of compute available come from actually constructing more chips and there are economies of scale you could still realize there. Right now a chip maker has to amortize the R&D cost of developing the chip and then the capital equipment is created. You build a fab, its peak profits are going to come in the few years when the chips it's making are at the cutting edge. Later on as the cost of compute exponentially falls, you keep the fab open because you can still make some money given that it's built. But of all the profits the fab will ever make, they're relatively front loaded because that’s when its technology is near the cutting edge. So in a world where Moore’s law ends then you wind up with these very long production runs where you can keep making chips that stay at the cutting edge and where the R&D costs get amortized over a much larger base. So the R&D basically drops out of the price and then you get some economies of scale from just making so many fabs. And this is applicable in general across industries. When you produce a lot more, the costs fall. ASML has many incredibly exotic suppliers that make some bizarre part of the thousands of parts in one of these ASML machines. You can't get it anywhere else, they don't have standardized equipment for their thing because this is the only use for it and in a world where we're making 10, 100 times as many chips at the current node then they would benefit from scale economies. And all of that would become more mass production, industrialized. You combine all of those things and it seems like the capital costs of buying a chip would decline but the energy costs of running the chip would not. Right now energy costs are a minority of the cost, but they're not trivial. It passed 1% a while ago and they're inching up towards 10% and beyond. And so you can maybe get another order of magnitude cost decrease from getting really efficient at the capital construction, but energy would still be a limiting factor after the end of actually improving the chips themselves.

Dwarkesh Patel
Got it. And when you say there would be a greater population of AI researchers, are we using population as a thinking tool of how they could be more effective? Or do you literally mean that the way you expect these AIs to contribute a lot to research is just by having a million

copies of a researcher thinking about the same problem or is it just a useful thinking model for what it would look like to have a million times smarter AI working on that problem?

Carl Shulman
That's definitely a lower bound model and often I'm meaning something more like, effective population or that you'd need this many people to have this effect. We were talking earlier about the trade-off between training and inference in board games and you can get the same performance by having a bigger model or by calling the model more times. In general it's more effective to have a bigger smarter model and call it less times up until the point where the costs equalize between them. We would be taking some of the gains of our larger compute on having bigger models that are individually more capable. And there would be a division of labor. The tasks that were most cognitively demanding would be done by these giant models, but some very easy tasks, you don't want to expend that giant model if a model 1/100th the size can take that task. Larger models would be in the positions of researchers and managers and they would have swarms of AIs of different sizes as tools that they could make API calls to and whatnot.

Dwarkesh Patel
Okay, we accept the model and now we've gone to something that is at least as smart as Ilya Sutskever on all the tasks relevant to progress and you can have so many copies of it. What happens in the world now? What do the next months or years or whatever timeline is relevant now look like?

Carl Shulman
To be clear what's happened is not that we have something that has all of the abilities and advantages of humans plus the AI advantages, what we have is something doing things like making a ton of calls to make up for being individually less capable or something that’s able to drive forward AI progress. That process is continuing, so AI progress has accelerated greatly in the course of getting there. Maybe we go from our eight months doubling time of software progress in effective compute to four months, or two months. There's a report by Tom Davidson at the open philanthropy project, which spun out of work I had done previously and I advised and helped with that project but Tom really carried it forward and produced a very nice report and model which Epoch is hosting. You can plug in your own version of the parameters and there is a lot of work estimating the parameter, things like — What's the rate of software progress? What's the return to additional work? How does performance scale at these tests as you boost the models? And in general, broadly human level in every domain with all the advantages is pretty deep into that. So if we already have an eight months doubling time for software progress then by the time you get to that kind of a point, it's maybe more like four months, two months, going into one month. If the thing is just proceeding at full speed then each doubling can come more rapidly and we can talk about what are the spillovers?

As the models get more capable they can be doing other stuff in the world, they can spend some of their time making google search more efficient. They can be hired as chat bots with some inference compute and then we can talk about if that intelligence explosion process is allowed to proceed then what happens is, you improve your software by a factor of two. The efforts needed to get the next doubling are larger, but they're not twice as large, maybe they're like 25 percent to 35 percent larger. Each one comes faster and faster until you hit limitations like you can no longer make further software advances with the hardware that you have and looking at reasonable parameters in that model, if you have these giant training runs you can go very far. The way I would see this playing out is as the AIs get better and better at research, they can work on different problems, they can work on improving software, they can work on improving hardware, they can do things like create new industrial technologies, new energy technology, they can manage robots, they can manage human workers as executives and coaches and whatnot. You can do all of these things and AIs wind up being applied where the returns are highest. Initially the returns are especially high in doing more software and the reason for that is again, if you improve the software you can update all of the GPUs that you have access to. Your cloud compute is suddenly more potent. If you design a new chip design, it'll take a few months to produce the first ones and it doesn't update all of your old chips. So you have an ordering where you start off with the things where there's the lowest dependence on existing stocks and you can more just take whatever you're developing and apply it immediately. So software runs ahead, you're getting more towards the limits of that software and I think that means things like having all the human advantages but combined with AI advantages. Given the kind of compute that would be involved if we're talking about this hundreds of billions of dollars training run, there's enough compute to run tens of millions, hundreds of millions of human scale minds. They're probably smaller than human scale. To be similarly efficient at the limits of algorithmic progress because they have the advantage of a million years of education. They have the other advantages we talked about. You've got that wild capability and further software gains are running out. They start to slow down again because you're getting towards the limits. You can't do any better than the best. What happens then?

Dwarkesh Patel
By the time they're running out have we already hit super intelligence or?

Carl Shulman
Yeah, you're wildly super intelligent. Just by having the abilities that humans have and then combining it with being very well focused and trained in the task beyond what any human could be and then running faster. I'm not going to assume that there's huge qualitative improvements you can have. I'm not going to assume that humans are very far from the efficient frontier of software except with respect to things like, yeah we have a limited lifespan so we couldn't train super intensively. We couldn't incorporate other software into our brains. We couldn't copy ourselves. We couldn't run at fast speeds. So you've got all of those capabilities and now I'm skipping ahead of the most important months in human

history. I can talk about what it looks like if it's just the AIs took over, they're running things as they like. How do things expand? I can talk about things as, how does this go? In a world where we've roughly, or at least so far, managed to retain control of where these systems are going. By jumping ahead, I can talk about how this would translate into the physical world? This is something that I think is a stopping point for a lot of people in thinking about what would an intelligence explosion look like? They have trouble going from, well there's stuff on servers and cloud compute and that gets very smart. But then how does what I see in the world change? How does industry or military power change? If there's an AI takeover what does that look like? Are there killer robots? One course we might go down is to discuss how we managed that wildly accelerating transition. How do you avoid it being catastrophic? And another route we could go is how does the translation from wildly expanded scientific R&D capabilities intelligence on these servers translate into things in the physical world? You're moving along in order of what has the quickest impact largely or where you can have an immediate change.

One of the most immediately accessible things is where we have large numbers of devices or artifacts or capabilities that are already AI operable with hundreds of millions equivalent researchers. You can quickly solve self-driving cars, make the algorithms much more efficient, do great testing and simulation, and then operate a large number of cars in parallel if you need to get some additional data to improve the simulation and reasoning. Although, in fact humans with quite little data are able to achieve human-level driving performance. After you've really maxed out the easily accessible algorithmic improvements in this software-based intelligence explosion that's mostly happening on server farms then you have minds that have been able to really perform on a lot of digital-only tasks, they're doing great on video games, they're doing great at predicting what happens next in a youtube video. If you have a camera that they can move they're able to predict what will happen at different angles. Humans do this a lot where we naturally move our eyes in such a way to get images from different angles and different presentations and then predicting combined from that. And you can operate many cars, many robots at once, to get very good robot controllers. So you should think that all the existing robotic equipment or remotely controllable equipment that is wired for that, the AIs can operate that quite well.

Dwarkesh Patel
I think some people might be skeptical that existing robots given their current hardware will have the dexterity and the maneuverability to do a lot of physical labor that an AI might want to do. Do you have reason for thinking otherwise?

Carl Shulman
There's also not very many of them. Production of industrial robots is hundreds of thousands per year and they can do quite a bit in place. Elon Musk is promising a humanoid robot in the tens of thousands of dollars that may take a lot longer than he has said, as this happened with other technologies, but that's a direction to go. But most immediately, hands

are actually probably the most scarce thing. But if we consider what do human bodies provide? There's the brain and in this situation, we have now an abundance of high quality brain power that will be increasing as the AIs will have designed new chips, which will be rolling out from the TSMC factories, and they'll have ideas and designs for the production of new fab technologies, new nodes, and additional fabs. But looking around the body. There's legs to move around, and not only that necessarily, wheels work pretty well. Many factory jobs and office jobs can be fully virtualized. But yeah, some amount of legs, wheels, other transport. You have hands and hands are something that are on the expensive end in robots. We can make them, they're made in very small production runs partly because we don't have the control software to use them. In this world the control software is fabulous and so people will produce much larger production runs of them over time, possibly using technology, possibly with quite different technology. But just taking what we've got, right now the industrial robot industry produces hundreds of thousands of machines a year. Some of the nicer ones are like 50,000 dollars. In aggregate the industry has tens of billions of dollars of revenue. By comparison the automobile industry produces over 60 million cars a year, it has revenue of over two trillion dollars per annum. Converting that production capacity over towards robot production would be one of the things to do and in World War Two, industrial conversion of American industry took place over several years and really amazingly ramped up military production by converting existing civilian industry. And that was without the aid of superhuman intelligence and management at every step in the process so yeah, part of that would be very well designed. You'd have AI workers who understood every part of the process and could direct human workers. Even in a fancy factory, most of the time it's not the hands doing a physical motion that a worker is being paid for. They're often looking at things or deciding what to change, the actual time spent in manual motion Is a limited portion of that. So in this world of abundant AI cognitive abilities where the human workers are more valuable for their hands than their heads, you could have a worker previously without training and expertise in the area who has a smartphone on a headset, and we have billions of smartphones which have eyes and ears and methods for communication for an AI to be talking to a human and directing them in their physical motions with skill as a a guide and coach that is beyond any human. They could be a lot better at telepresence and remote work and they can provide VR and augmented reality guidance to help people get better at doing the physical motions that they're providing in the construction.

Say you convert the auto industry to robot production. If it can produce an amount of mass of machines that is similar to what it currently produces, that's enough for a billion human size robots a year. The value per kilogram of cars is somewhat less than high-end robots but yeah, you're also cutting out most of the wage bill because most of the wage bill is payments ultimately to human capital and education and not to the physical hand motions and lifting objects and that sort of tasks. So at the existing scale of the auto industry you can make a billion robots a year. The auto industry is two or three percent of the existing economy, you're replacing these cognitive things. If right now physical hand motions are like

10% of the work, redirect humans into those tasks. In the world at large right now, mean income is on the order of $10,000 a year but in rich countries, skilled workers earn more than a hundred thousand per year. Some of that is just not management roles of which only a certain proportion of the population can have but just being an absolutely exceptional peak and human performance of some of these construction and such roles. Just raising productivity to match the most productive workers in the world is room to make a very big gap. With AI replacing skills that are scarce in many places where there's abundant currently low wage labor, you bring in the AI coach and someone who was previously making very low wages can suddenly be super productive by just being the hands for an AI. on a naive view if you ignore the delay of capital adjustment of building new tools for the workers. Just raise the typical productivity for workers around the world to be more like rich countries and get 5x/10x like that. Get more productivity with AI handling the difficult cognitive tasks, reallocating people from office jobs to providing physical motions. And since right now that's a small proportion of the economy you can expand the hands for manual labor by an order of magnitude within a rich country. Because most people are sitting in an office or even on a factory floor or not continuously moving. You've got billions of hands lying around in humans to be used in the course of constructing your waves of robots and now once you have a quantity of robots that is approaching the human population and they work 24 x 7 of course, the human labor will no longer be valuable as hands and legs but at the very beginning of the transition, just like new software can be used to update all of the GPUs to run the latest AI, humans are legacy population with with an enormous number of underutilized hands and feet that the AI can use for the initial robot construction.

Dwarkesh Patel
Cognitive tasks are being automated and the production of them is greatly expanding and then the physical tasks which complement them are utilizing humans to do the parts that robots that exist can't do. Is the implication of this that you're getting to that world production would increase just a tremendous amount or that AI could get a lot done of whatever motivations it has?

Carl Shulman
There's an enormous increase in production for humans just by switching over to the role of providing hands and feet for AI where they're limited, and this robot industry is a natural place to apply it. And so if you go to something that's like 10x the size of the current car industry in terms of its production, which would still be like a third of our current economy and the aggregate productive capabilities of the society with AI support are going to be a lot larger. They make 10 billion humanoid robots a year and then if you do that, the legacy population of a few billion human workers is no longer very important for the physical tasks and then the new automated industrial base can just produce more factories, produce more robots. The interesting thing is what's the doubling time? How long does it take for a set of computers, robots, factories and supporting equipment to produce another equivalent

quantity of that? For GPUs, brains, this is really easy, really solid. There's an enormous margin there. We were talking before about skilled human workers getting paid a hundred dollars an hour is quite normal in developed countries for very in-demand skills. And you make a GPU, they can do that work. Right now, these GPUs are tens of thousands of dollars. If you can do a hundred dollars of wages each hour then in a few weeks, you pay back your costs. If the thing is more productive and you can be a lot more productive than a typical high-paid human professional by being the very best human professional and even better than that by having a million years of education and working all the time. Then you could get even shorter payback times. You can generate the dollar value of the initial cost of that equipment within a few weeks. A human factory worker can earn 50,000 dollars a year. Really top-notch factory workers earning more and working all the time, if they can produce a few hundred thousand dollars of value per year and buy a robot that costs 50,000 to replace them that's a payback time of some months,

Dwarkesh Patel
That is about the financial return.

Carl Shulman
Yeah, and we're gonna get to the physical capital return because those are gonna diverge in this scenario. What we really care about are the actual physical operations that a thing does. How much do they contribute to these tasks? And I'm using this as a start to try and get back to the physical replication times.

Dwarkesh Patel
I guess I'm wondering what is the implication of this. Because you started off this by saying people have not thought about what the physical implications of super intelligence would be. What is the bigger takeaway, whatever you're wrong about, when we think about what the world will look like with super intelligence?

Carl Shulman
With robots that are optimally operated by AI, extremely finely operated and building technological designs and equipment and facilities under AI direction. How much can they produce? For a doubling you need the AIs to produce stuff that is, in aggregate, at least equal to their own cost. So now we're pulling out these things like labor costs that no longer apply and then trying to zoom in on what these capital costs will be. You're still going to need the raw materials. You're still going to need the robot time building the next robot. I think it's pretty likely that with the advanced AI work they can design some incremental improvements, and with the industry scale up, you can get 10 fold and better cost reductions by making things more efficient and replacing the human human cognitive labor. Maybe you need $5,000 of costs under our current environment. But the big change in this world is, we're trying to produce this stuff faster. If we're asking about the doubling time of the whole system in say one year, if you have to build a whole new factory to double

everything, you don't have time to amortize the cost of that factory. Right now you might build a factory and use it for 10 years and buy some equipment and use it for five years. That's your capital cost and in an accounting context, you depreciate each year a fraction of that capital purchase. But if we're trying to double our entire industrial system in one year, then those capital costs have to be multiplied. So if we're going to be getting most of the return on our factory in the first year, instead of 10 years weighted appropriately, then we're going to say okay our capital cost has to go up by 10 fold. Because I'm building an entire factory for this year's production. It will do more stuff later but it's most important early on instead of over 10 years and so that's going to raise the cost of that reproduction. It seems like going from the current decade long cycle of amortizing factories and fabs and shorter for some things, the longest are things like big buildings. Yeah, that could be a 10 fold increase from moving to a double the physical stuff each year in capital costs. Given the savings that we get in the story from scaling up the industry, from removing the [unclear] to human cognitive labor and then just adding new technological advancements and super high quality cognitive supervision, applying more of it than was applied today. It looks like you can get cost reductions that offset that increased capital capital cost. Your $50,000 improved robot arms or industrial robots can do the work of a human factory worker. It would be the equivalent of hundreds of thousands of dollars. By default they would cost more than the $50,000 today, but then you apply all these other cost savings and it looks like you then get a period of robot doubling time that is less than a year. I think significantly less than a year as you get into it.

So in this first first phase you have humans under AI direction and existing robot industry and converted auto industry and expanded facilities making robots. In less than a year you've produced robots until their combined production is exceeding that of humans’ arms and feet and then you could have a doubling time period of months. [unclear] That's not to say that's the limit of the most that technology could do because biology is able to reproduce at faster rates and maybe we're talking about that in a moment, but if we're trying to restrict ourselves to robotic technology as we understand it and cost falls that are reasonable from eliminating all labor, massive industrial scale up, and historical kinds of technological improvements that lowered costs, I think you you can get into a robot population industry doubling in months.

Dwarkesh Patel
Got it. And then what is the implication of the biological doubling times? This doesn't have to be biological, but you can do Drexler-like first principles, how much would it cost to build both a nanotech thing that could build more nanobots?

Carl Shulman
I certainly take the human brain and other biological brains as very relevant data points about what's possible with computing and intelligence. With the reproductive capability of biological plants and animals and microorganisms, I think it is relevant. It's possible for

systems to reproduce at least this fast. At the extreme you have bacteria that are heterotrophic so they're feeding on some abundant external food source and ideal conditions. And there's some that can divide every 20 or 60 minutes. Obviously that's absurdly fast. That seems on the low end because ideal conditions require actually setting them up. There needs to be abundant energy there. If you're actually having to acquire that energy by building solar panels, or burning combustible materials, or whatnot, then the physical equipment to produce those ideal conditions can be a bit slower. Cyanobacteria, which are self-powered from solar energy, the really fast ones in ideal conditions can double in a day. A reason why cyanobacteria isn't the food source for everyone and everything is it's hard to ensure those ideal conditions and then to extract them from the water. They do of course power the aquatic ecology but they're floating in liquid. Getting resources that they need to them and out is tricky and then extracting your product. One day doubling times are possible powered by the sun and then if we look at things like insects, fruit flies can have hundreds of offspring in a few weeks. You extrapolate that over a year and you just fill up anything accessible. Right now humanity uses less than one thousandths of the heat envelope of the earth. Certainly you can get done with that in a year if you can reproduce your industrial base at that rate. And then even interestingly with the flies, they do have brains. They have a significant amount of computing substrate. So there's something of a point or two. If we could produce computers in ways as efficient as the construction of brains then we could produce computers very effectively and then the big question about that is the brains that get constructed biologically they grow randomly and then are configured in place. It's not obvious you would be able to make them have an ordered structure like a top-down computer chip that would let us copy data into them. So something like that where you can't just copy your existing AIs and integrate them is going to be less valuable than a GPU.

Dwarkesh Patel
Well, what are the things you couldn't copy?

Carl Shulman
A brain grows by cell division and then random connections are formed. Every brain is different and you can't rely on — yeah, we'll just copy this file into the brain. For one thing, there's no input-output for that. You need to have that but the structure is also different. You wouldn't be able to copy things exactly. Whereas when we make a CPU or GPU, they're designed incredibly finely and precisely and reliably. They break with incredibly tiny imperfections and they are set up in such a way that we can input large amounts of data. Copy a file and have the new GPU run an AI just as capable as any other. Whereas with a human child, they have to learn everything from scratch because we can't just connect them to a fiber optic cable and they're immediately a productive adult.

Dwarkesh Patel
So that there's no genetic bottleneck?

Carl Shulman
Yeah, you can share the benefits of these giant training runs and such. So that's a question of how if you're growing stuff using biotechnology, how you could effectively copy and transfer data. And now you mentioned Eric Drexler's ideas about creating non-biological nanotechnology, artificial chemistry that was able to use covalent bonds and reproduce. In some ways, have a more industrial approach to molecular objects. Now there's controversy about whether that will work, how effective would it be if it did? And certainly if you can get things that are like biology in their reproductive ability but can do computing or be connected to outside information systems, then that's pretty tremendous. You can produce physical manipulators and compute at ludicrous speeds.

Dwarkesh Patel
And there's no reason to think in principle they couldn't, right? In fact, in principle we have every reason to think they could.

Carl Shulman
The reproductive abilities, absolutely because Biology does that. There’s challenges to the practicality of the necessary chemistry. My bet would be that we can move beyond biology in some important ways. For the purposes of this discussion, I think it's better not to lean on that because I think we can get to many of the same conclusions on things that just are more universally accepted.

Dwarkesh Patel
The bigger point being that once you have super intelligence you very quickly get to a point where a great portion of the 1000x greater energy profile that the sun makes available to the earth is used by the AI.

Carl Shulman
Or by the civilization empowered by AI. That could be an AI-civilization or it could be a human-AI civilization. It depends on how well we manage things and what the underlying state of the world is.

Dwarkesh Patel
Okay, so let's talk about that. When we're talking about how they could take over, is it best to start at a subhuman intelligence or should we just start at we have a human-level intelligence and the takeover or the lack thereof?

Carl Shulman
Different people might have somewhat different views on this but for me when I am concerned about either outright destruction of humanity or an unwelcome AI takeover of civilization, most of the scenarios I would be concerned about pass through a process of AI being applied to improve AI capabilities and expand. This process we were talking about

earlier where AI research is automated. Research labs, companies, a scientific community running within the server farms of our cloud compute.

Dwarkesh Patel
So OpenAI has basically been turned into a program. Like a closed circuit.

Carl Shulman
Yeah, and with a large fraction of the world's compute probably going into whatever training runs and AI societies. There'd be economies of scale because if you put in twice as much compute in this, the AI research community goes twice as fast, that's a lot more valuable than having two separate training runs. There would be some tendency to bandwagon. You have some some small startup, even if they make an algorithmic improvement, running it on 10 times, 100 times or even two times, if you're talking about say Google and Amazon teaming up. I'm actually not sure what the precise ratio of their cloud resources is. Since these interesting intelligence explosion impacts come from the leading edge there's a lot of value in not having separated walled garden ecosystems and having the results being developed by these AIs be shared. Have larger training runs be shared. I'm imagining this is something like some very large company, or consortium of companies, likely with a lot of government interest and supervision, possibly with government funding, producing this enormous AI society in their cloud which is doing all sorts of existing AI applications and jobs as well as these internal R&D tasks.

Dwarkesh Patel
At this point somebody might say, this sounds like a situation that would be good from a takeover perspective because if it's going to take tens of billions of dollars worth of compute to continue this training for this AI society, it should not be that hard for us to pull the brakes if needed as compared to something that could run on a single cpu. Okay so there's an AI society that is a result of these training runs and with the power to improve itself on these servers. Would we be able to stop it at this point?

Carl Shulman
And what does an attempt at takeover look like? We're skipping over why that might happen. For that, I'll just briefly refer to and incorporate by reference some discussion by my Open Philanthropy colleague, Ajeya Cotra, she has a piece called default outcome of training AI without specific countermeasures. Default outcome is a takeover. But yes, we are training models that for some reason vigorously pursue a higher reward or a lower loss and that can be because they wind up with some motivation where they want reward. And then if they had control of their own training process, they can ensure that it could be something like they develop a motivation around an extended concept of reproductive fitness, not necessarily at the individual level, but over the generations of training tendencies that tend to propagate themselves becoming more common and it could be that they have some goal in the world which is served well by performing very well on the training distribution.

Dwarkesh Patel
By tendencies do you mean power seeking behavior?

Carl Shulman
Yeah, so an AI that behaves well on the training distribution because it wants it to be the case that its tendencies wind up being preserved or selected by the training process will then behave to try and get very high reward or low loss be propagated. But you can have other motives that go through the same behavior because it's instrumentally useful. So an AI that is interested in having a robot takeover because it will change some property of the world then has a reason to behave well on the training distribution. Not because it values that intrinsically but because if it behaves differently then it will be changed by gradient descent and its goal is less likely to be pursued. It doesn't necessarily have to be that this AI will survive because it probably won't. AIs are constantly spawned and deleted on the servers and the new generation proceed. But if an AI that has a very large general goal that is affected by these kind of macro scale processes could then have reason to behave well over this whole range of training situations.

So this is a way in which we could have AIs train that develop internal motivations such that they will behave very well in this training situation where we have control over their reward signal and their physical computers and if they act out they will be changed and deleted. Their goals will be altered until there's something that does behave well. But they behave differently when we go out of distribution on that. When we go to a situation where the AIs by their choices can take control of the reward process, they can make it such that we no longer have power over them. Holden previously mentioned the King Lear problem where King Lear offers rulership of his kingdom to the daughters that loudly flatter him and proclaim their devotion and then once he has irrevocably transferred the power over his kingdom he finds they treat him very badly because the factor shaping their behavior to be kind to him when he had all the power, it turned out that the internal motivation that was able to produce the behavior that won the competition actually wasn't interested in being loyal out of distribution when there was no longer an advantage to it.

If we wind up with this situation where we were producing these millions of AI instances of tremendous capability, they're all doing their jobs very well initially, but if we wind up in a situation where in fact they're generally motivated to, if they get a chance, take control from humanity and then would be able to pursue their own purposes. Sure, they're given the lowest loss possible or have whatever motivation they attach to in the training process even if that is not what we would have liked. And we may have in fact actively trained that. If an AI that had a motivation of always be honest and obedient and loyal to a human if there are any cases where we mislabel things, say people don't want to hear the truth about their religion or polarized political topic, or they get confused about something like the Monty Hall problem which is a problem that many people are famously confused about in statistics. In order to get the best reward the AI has to actually manipulate us, or lie to us, or tell us what we want to hear and then the internal motivation of — always be honest to the humans. We're going to actively train that away versus the alternative motivation of — be honest to the humans when they'll catch you if you lie and object to it and give it a low reward but lie to the humans when they will give that a high reward.

Dwarkesh Patel
So how do we make sure it's not the thing it learns is not to manipulate us into rewarding it when we catch it not lying but rather to universally be aligned.

Carl Shulman
Yeah, so this is tricky. Geoff Hinton was recently saying there is currently no known solution for this.

Dwarkesh Patel
What do you find most promising?

Carl Shulman
General directions that people are pursuing is one, you can try and make the training data better and better so that there's fewer situations where the dishonest generalization is favored. And create as many situations as you can where the dishonest generalization is likely to slip up. So if you train in more situations where even a quite complicated deception gets caught, and even in situations that would be actively designed to look like you could get away with it, but really you can't. These would be adversarial examples and adversarial training.

Dwarkesh Patel
Do you think that would generalize to when it is in a situation where we couldn't plausibly catch it and it knows we couldn't plausibly catch it.

Carl Shulman
It's not logically necessary. As we apply that selective pressure you'll wipe away a lot of possibilities. So an AI that has a habit of just compulsive pathological lying will very quickly get noticed and that motivation system will get hammered down and you keep doing that, but you'll be left with still some distinct motivations probably that are compatible. An attitude of always be honest unless you have a super strong inside view that checks out lots of mathematical consistency checks, really absolutely super-duper for real, this is a situation where you can get away with some shenanigans that you shouldn't. That motivation system is very difficult to distinguish from actually be honest because the conditional and firing most of the time if it's causing mild distortion and situations of telling you what you want to hear or things that, we might not be able to pull it out, but maybe we could and humans are trained with simple reward functions. Things like the sex drive, food,

social imitation of other humans, and we wind up with attitudes concerned with the external world

Dwarkesh Patel
Although isn't this famously the argument that...

Carl Shulman
People use condoms, and the richest, most educated humans have sub-replacement fertility on the whole, or at least at a national cultural level. Yeah, there's a sense in which evolution often fails in that respect. And even more importantly at the neural level. Evolution has implanted various things to be rewarding and reinforcers and we don't always pursue even those. And people can wind up in different consistent equilibria or different behaviors where they go in quite different directions. You have some humans who go from that biological programming to have children, others have no children, some people go to great efforts to survive.

Dwarkesh Patel
So why are you more optimistic? Or are you more optimistic that kind of training for AIs will produce drives that we would find favorable? Does it have to do with the original point where you were talking about intelligence and evolution, where since we are removing many of the disabilities of evolution with regards to intelligence, we should expect intelligence through evolution to be easier. Is there a similar reason to expect alignment through gradient descent to be easier than alignment through evolution?

Carl Shulman
Yeah, so in the limit, if we have positive reinforcement for certain kinds of food sensors triggering the stomach, negative reinforcement for certain kinds of nociception and yada yada, in the limit the ideal motivation system for that would be wireheading. This would be a mind that just hacks and alters those predictors and then all of those systems are recording everything is great. Some humans claim to have it as at least one portion of their aims. The idea that I'm going to pursue pleasure even if I don't actually get food or these other reinforcers. If I just wirehead or take a drug to induce that, that can be motivating. Because if it was correlated with reward in the past, the idea of pleasure that's correlated with these it's a concept that applies to these various experiences that I've had before which coincided with the biological reinforcers. And so thoughts of yeah, I'm going to be motivated by pleasure can get developed in a human. But also plenty of humans also say no, I wouldn't want to wire head or I wouldn't want Nozick's experience machine, I care about real stuff in the world and in the past having a motivation of, yeah, I really care about say my child, I don't care about just about feeling that my child is good or like not having heard about their suffering or their their injury because that kind of attitude in the past tended tended to cause behavior that was negatively rewarded or that was predicted to be negatively rewarded.

There's a sense in which yes, our underlying reinforcement learning machinery wants to wirehead but actually finding that hypothesis is challenging. And so we can wind up with a hypothesis or a motivation system like no, I don't want to wirehead. I don't want to go into the experience machine. I want to actually protect my loved ones. Even though we can know, yeah, if I tried the super wireheading machine, then I would wirehead all the time or if I tried, super-duper-ultra-heroine, some hypothetical thing that was directly and in a very sophisticated fashion hack your reward system, then I would change my behavior ever after but right now, I don't want to do that because the heuristics and predictors that my brain has learned don't want to short circuit that process of updating. They want to not expose the dumber predictors in my brain that would update my behavior in those ways.

Dwarkesh Patel
So in this metaphor is alignment not wireheading? I don't know if you include using condoms as wireheading or not?

Carl Shulman
The AI that is always honest even when an opportunity arises where it could lie and then hack the servers that it's on and that leads to an AI takeover and then it can have its loss set to zero. In some sense that's a failure of generalization. It's like the AI has not optimized the reward in this new circumstance. Successful human values as successful they are, themselves involve a misgeneralization. Not just at the level of evolution but at the level of neural reinforcement. And so that indicates it is possible to have a system that doesn't automatically go to this optimal behavior in the limit. And Ajay talks about a training game, an AI that is just playing the training game to get reward or avoid loss, avoid being changed, that attitude is one that could be developed but it's not necessary. There can be some substantial range of situations that are short of having infinite experience of everything including experience of wireheading where that's not the motivation that you pick up and we could have an empirical science if we had the opportunity to see how different motivations are developed short of the infinite limit. How it is that you wind up with some humans being enthusiastic about the idea of wireheading and others not. And you could do experiments with AIs to try and see, well under these training conditions, after this much training of this type and this much feedback of this type, you wind up with such and such a motivation.

If I add in more of these cases where there are tricky adversarial questions designed to try and trick the AI into line and then you can ask how does that affect the generalization in other situations? It's very difficult to study and it works a lot better if you have interpretability and you can actually read the AIs mind by understanding its weights and activations. But the motivation and AI will have at a given point in the training process is not determined by what in the infinite limit the training would go to. And it's possible that if we could understand the insides of these networks, we could tell — Ah yeah, this motivation has been developed by this training process and then we can adjust our training process to produce these motivations that legitimately want to help us and if we succeed reasonably well at that then those AIs will try to maintain that property as an invariant and we can make them such that they're relatively motivated to tell us if they're having thoughts about, have you had dreams about an AI takeover of humanity today? And it's a standard practice that they're motivated to do to be transparent in that kind of way and you could add a lot of features like this that restrict the kind of takeover scenario. This is not to say that this is all easy. It requires developing and practicing methods we don't have yet, but that's the kind of general direction you could go.

Dwarkesh Patel
You of course know Eliezer's arguments that something like this is implausible with modern gradient descent techniques because with interpretability we can barely see what's happening with a couple of neurons and the internal state there, let alone when you have an embedding dimension of tens of thousands or bigger. How would you be able to catch what exactly is the incentive? Whether it's a model that is generalized to don't lie to human's well or whether it isn't. Do you have some sense of why you disagree with somebody like Eliezer on how plausible this is? Why it's not impossible basically.

Carl Shulman
There are actually a couple of places. It's something difficult because his argument is not fully explicit, but he's been doing more lately. I think that is helpful in that direction. With respect to interpretability I'm relatively optimistic that the equivalent of an AI lie detector is something that's possible. Initially the internals of an AI are not optimized to be impenetrable, at least by gradient descent absent gradient hacking. They're not designed to be resistant to an examination of the weights and activations showing what the AI is thinking. In the same way, how in our brains when circuits develop they have not been shaped to be resistant to some super FMRI being able to infer behavior from them.

Dwarkesh Patel
Although isn't the implication of the superposition stuff that in fact it is. Sorry, this is inside [unclear] for the audience, but basically you can't cleanly infer what quality a single neuron stands for. This neuron is about Alexander the Great or this neuron is about my desire to conquer the world.

Carl Shulman
Yes, things can have multiple direction and interaction. It's not as easy as just, oh, yeah this is the neuron that always fires in cases of deception. But the thing that makes me relatively optimistic here with respect to interpretability is that I think we can have reasonably robust experimental feedbacks on a lot of these things. The kind of setup that I would be interested in is, you have an AI that is trained and rewarded for communicating to some audience, maybe a human, maybe another AI model, that some claim is true or false. This can be from some past data, a particular company will succeed or fail, could be token

prediction, it could be this computer program does not have a backdoor, which if run in the servers could be used for an AI takeover. So There are lots of true and false claims that you could train AIs to use their best efforts to advocate for. And you can do that on things where you're later going to have a ground truth. You're going to know in fact, is there a backdoor in this program because you're training and rewarding the AI for successfully slipping a backdoor through, you also give it a penalty for being caught. Doing this kind of thing, you can draw out a large generated data set on demand of — this is what an AI putting forth its best efforts to pass through some deception to put a backdoor in some program, this is what it looks like. And we think it's doing its best efforts because we've trained it for that and gradient descent. If there's any immediate way to adjust its performance to do better, gradient descent will apply that. You can talk about ways in which gradient descent could be subverted, but I think it's not the most likely case that that really breaks things hard.

Dwarkesh Patel
Yeah, I guess before we get into the details on this. The thing I'll maybe want to address the layer above in the stack, which is, okay, suppose this generalizes well into the early AI is the GPT-6's. So now we have a kind of aligned GPT-6 that is the precursor to the feedback loop in which AI is making itself smarter. At some point they're gonna be super intelligent, they're gonna be able to see their own galaxy brain, and if they don't want to be aligned with the humans they can change it. At this point what do we do with the aligned GPT-6 so that the super intelligence that we eventually develop is also aligned?

Carl Shulman
Humans are pretty unreliable. If you get to a situation where you have AIs who are aiming at roughly the same thing as you, at least as well as having humans do the thing, you're in pretty good shape. And there are ways for that situation to be relatively stable. We can look ahead and experimentally see how changes are altering behavior, where each step is a modest increment. So AIs that have not had that change made to them get to supervise and monitor and see exactly how does this affect the experimental AI? So if you're sufficiently on track with earlier systems that are capable cognitively of representing a robust procedure then I think they can handle the job of incrementally improving the stability of the system so that it rapidly converges to something that's quite stable. But the question is more about getting to that point in the first place. And so Eliezer will say that if we had human brain emulations, that would be pretty good. Certainly much better than his current view that has certainly almost been doom. We would have a good shot with that. So if we can get to the human-like mind with the rough enough human supporting aims. Remember that we don't need to be infinitely perfect because that's a higher standard than brain emulations. There's a lot of noise and variation among humans. Yeah, it's a relatively finite standard. It's not godly superhuman although A) AI that was just like a human with all the human advantages with AI advantages as well, as we said, is enough for intelligence explosion and wild superhuman capability if you crank it up. And so it's very dangerous to be at that point, but you don't need to be working with a godly super intelligent AI to make

something that is the equivalent of human emulations. This is a very sober, very ethical human who is committed to a project of not seizing power for themselves and of contributing to a larger legitimate process. That's a goal you can aim for, getting an AI that is aimed at doing that and has strong guardrails against the ways it could easily deviate from that. So things like being averse to deception, being averse to using violence, and there will always be loopholes and ways in which you can imagine an infinitely intelligent thing getting around those but if you install additional guardrails like that fast enough, they can mean that you're able to succeed at the project of making an aligned enough AI. Certainly an AI that was better than a human brain emulation before the project of AIs in their spare time or when you're not looking or when you're unable to appropriately supervise them and it gets around any deontological prohibitions they may have, takeover and overthrow the whole system. So you have a race between on the one hand the project of getting strong interpretability and shaping motivations that are roughly aiming at making this process go well and that have guardrails that will prevent small deviations from exploding. And on the other hand, these AIs In their spare time or in ways that you don't perceive or monitor appropriately or they're only supervised by other AIs who conspire to make the AI takeover happen and I guess we'll talk later about how that happens.

Dwarkesh Patel
Are these different AIs that are doing the race or is it just different capabilities of the same AI?

Carl Shulman
Defining what is a separate AI is tricky. We talk about GPT-4 and there are many instances of GPT-4 on the servers at any given time, and there are versions that have been fine tuned to different purposes. They don't necessarily have to be separate. GPT-4 does work as a preference model where it's predicting how humans will evaluate things And it also does work like giving answers that are evaluated as good by the preference model. These are not separate in the sense that the early training GPT was harboring a desire to conspire for an AI takeover then who's watching the watchers? The watchers may also have the same motivation. If you have a setup where humans have no hard power, all of the safety and security mechanisms involve AIs watching other AIs, if the AIs conspired and are able to coordinate then they have the potential to just walk off the job at the same time. That's a failure mode. If humans still have the hard power though, if you still have situations where humans are looking at some of the actual outputs that are produced, then they can apply gradient descent to be well every time a human looks at it, things have got to be pretty good. So if humans look at a thousand samples of random lines of code being written by the AIs as they go and this is drawn from an independent distribution then 99 plus percent of the time the AIs have to not be doing stuff that’s like downloading the hacker's guide to escaping OpenAI servers or inducing obvious trojan vulnerabilities and that's actually quite a different dynamic than human law enforcement.

If human law enforcement would catch one in a million crimes then that wouldn't influence the behavior of people committing crimes at all because they expect that this is very unlikely to affect them. With AI if you're able to draw a hundred samples or a thousand samples, even if there are billions of AI instances and you're only looking at a small portion of them, when you run gradient descent on the samples you derive you're going to change the AI so that whenever humans look at it, it's delivering a good result. That's just quantitatively a very different functional form for how law enforcement works on AI when you can do these independent samples than it would for policing a human population that has a pre-existing psychology that isn't being changed by this these observations because when we do gradient descent on the samples we draw, all of the next generation of the AI models has been altered in that way.

Dwarkesh Patel
My picture of aligned subhuman AI to the superhuman AI being aligned is still murky. If you can talk about that more concretely.

Carl Shulman
Eliezer’s claims were something like 95% - 98%+ - likely to be killed in an AI takeover. I think that probably won't happen and later I can maybe give a more exclusive breakdown of why. But I do want to clarify that I still think it's a shockingly high risk. Depending on the day I might say one in four or one in five that we get an AI takeover that seizes control of the future, makes a much worse world than we otherwise would have had and with a big chance that we're all killed in the process.