Dwarkesh Podcast #62 - Paul Christiano - Preventing an AI Takeover

Published - October 31, 2023

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Okay, today I have the pleasure of interviewing Paul Christiano, who is the leading AI safety researcher. He's the person that labs and governments turn to when they want feedback and advice on their safety plans. He previously led the Language Model Alignment team at OpenAI, where he led the invention of RLHF. And now he is the head of the Alignment Research Center. And they've been working with the big labs to identify when these models will be too unsafe to keep scaling. Paul, welcome to the podcast.

Paul Christiano
Thanks for having me. Looking forward to talking.

Dwarkesh Patel
Okay, so first question. And this is a question I've asked Holden, Ilya, Dario, and none of them are going to be a satisfying answer. Give me a concrete sense of what a post-AGI world that would be good would look like. How are humans interfacing with the AI? What is the economic and political structure?

Paul Christiano
Yeah, I guess this is a tough question for a bunch of reasons. Maybe the biggest one is concrete. And I think it's just if we're talking about really long spans of time, then a lot will change. And it's really hard for someone to talk concretely about what that will look like without saying really silly things. But I can venture some guesses or fill in some parts. I think this is also a question of how good is good? Like, often I'm thinking about worlds that seem like kind of the best achievable outcome or a likely achievable outcome. So I am very often imagining my typical future has sort of continuing economic and military competition amongst groups of humans. I think that competition is increasingly mediated by AI systems. So, for example, if you imagine humans making money, it'll be less and less worthwhile for humans to spend any of their time trying to make money or any of their time trying to fight wars. So increasingly, the world you imagine is one where AI systems are doing those activities on behalf of humans. So, like, I just invest in some index fund, and a bunch of AIS are running companies, and those companies are competing with each other. But that is kind of a sphere where humans are not really engaging much. The reason I gave this how good is good caveat is, like, it's not clear if this is the world you'd most love. I'm like, yeah, I'm leading with like, the world still has a lot of war and of economic competition and so on. But maybe what I'm trying to what I'm most often thinking about is, like, how can a world be reasonably good during a long period where those things still exist? In the very long run, I kind of expect something more like strong world government rather than just this status quo. That's like, a very long run. I think there's, like, a long time left of having a bunch of states and a bunch of different economic powers, one world government.

Dwarkesh Patel
Why do you think that's the transition that's likely to happen at some point.

Paul Christiano
So again, at some point I'm imagining, or I'm thinking of the very broad sweep of history. I think there are a lot of losses. Like war is a very costly thing. We would all like to have fewer wars. If you just ask what is humanity's long term future like? I do expect to drive down the rate of war to very, very low levels eventually. It's sort of like this kind of technological or sociotechnological problem of sort of how do you organize society, navigate conflicts in a way that doesn't have those kinds of losses. And in the long run, I do expect this to succeed. I expect it to take kind of a long time. Subjectively, I think an important fact about AI is just like doing a lot of cognitive work and more quickly, getting you to that world more quickly, or figuring out how do we set things up that way?

Dwarkesh Patel
Yeah, the way Carl Schulman put it on the podcast is that you would have basically a thousand years of intellectual progress or social progress in a span of a month or whatever when the intelligence explosion happens more broadly. So the situation know we have these AIS who are managing our hedge funds and managing our factories and so on. That seems like something that makes sense when the AI is human level. But when we have superhuman AIS, do we want gods who are enslaved forever in 100 years? What is the decision we want?

Paul Christiano
100 years is a very, very long time. Maybe starting with the spirit of the question. Or maybe I have a view which is perhaps less extreme than Carl's view, but still like a hundred objective years is further ahead than I ever think. I still think I'm describing a world which involves incredibly smart systems running around, doing things like running companies on behalf of humans and fighting wars on behalf of humans. And you might be like, is that the world you really want? Or certainly not the first best world, as we mentioned a little bit before, I think it is a world that probably is of the achievable worlds or like feasible worlds is the one that seems most desirable to me that is sort of decoupling the social transition from this technological transition. So you could say, like, we're about to build some AI systems, and at the time we build AI systems, you would like to have either greatly changed the way world government works, or you would like to have sort of humans have decided like, we're done, we're passing off the baton to these AI systems. I think that you would like to decouple those timescales. So I think AI development is by default, barring some kind of coordination going to be very fast. So there's not going to be a lot of time for humans to think like, hey, what do we want? If we're building the next generation instead of just raising it the normal way. Like, what do we want that to look like? I think that's like a crazy hard kind of collective decision that humans naturally want to cope with over a bunch of generations patients. And the construction of AI is this very fast technological process happening over years. So I don't think you want to say like, by the time we have finished this technological progress, we will have made a decision about the next species we're going to build and replace ourselves with. I think the world we want to be in is one where we say either we are able to build the

technology in a way that doesn't force us to have made those decisions, which probably means it's a kind of AI. System that we're happy, like Delegating fighting a war, running a company to, or if we're not able to do that, then I really think you should not be doing you shouldn't have been building that technology. If you're like, the only way you can cope with AI is being ready to hand off the world to some AI system you built. I think it's very unlikely we're going to be sort of ready to do that. On the timelines that the technology would naturally dictate, say we're in the situation.

Dwarkesh Patel
In which we're happy with the thing. What would it look like for us to say we're ready to hand off the baton? What would make you satisfied? And the reason it's relevant to ask you is because you're on Anthropic's Long-Term Benefit Trust and you'll choose the majority of the board members. In the long run at Anthropic, these will presumably be the people who decide if Anthropic gets AI first, what the AI ends up doing. So what is the version of that that you would be happy with?

Paul Christiano
My main high level take here is that I would be unhappy about a world where Anthropic just makes some call and Anthropic is like, here's the kind of AI. We've seen enough, we're ready to hand off the future to this kind of AI. So procedurally, I think it's not a decision that kind of I want to be making personally or I want Anthropic to be making. So I kind of think from the perspective of that decision making are those challenges? The answer is pretty much always going to be like, we are not collectively ready because we're sort of not even all collectively engaged in this process. And I think from the perspective of an AI company, you kind of don't have this fast handoff option. You kind of have to be doing the option value to build the technology in a way that doesn't lock humanity into one course path. This isn't answering your full question, but this is answering the part that I think is most relevant to governance questions for Anthropic.

Dwarkesh Patel
You don't have to speak on behalf of Anthropic. I'm not asking about the process by which we would, as a civilization, agree to hand off. I'm just saying, okay, I personally, it's hard for me to imagine in 100 years that these things are still our slaves. And if they are, I think that's not the best world. So at some point, we're handing off the baton. Where would you be satisfied with this is an arrangement between the humans and AIS where I'm happy to let the rest of the universe or the rest of time play out.

Paul Christiano
I think that it is unlikely that in 100 years I would be happy with anything that was like, you had some humans, you're just going to throw away the humans and start afresh with these machines you built. That is I think you probably need subjectively longer than that before I or most people are like, okay, we understand what's up for grabs here. If you talk about 100

years, I kind of do. There's a process that I kind of understand and like a process of like, you have some humans. The humans are, like, talking and thinking and deliberating together. The humans are having kids and raising kids, and one generation comes after the next. There's that process we kind of understand, and we have a lot of views about what makes it go well or poorly, and we can try and improve that process and have the next generation do it better than the previous generation. I think there's some story like that that I get and that I like. And then I think that the default path to be comfortable with something very different is kind of more like just run that story for a long time, have more time for humans to sit around and think a lot and conclude, here's what we actually want. Or a long time for us to talk to each other or to grow up with this new technology and live in that world for our whole lives and so on. And so I'm mostly thinking from the perspective of these more local changes of saying not like, what is the world that I want? What's the crazy world? The kind of crazy I'd be happy handing off to more, just like, in what way do I wish we right now were different? How could we all be a little bit better? And then if we were a little bit better, then they would ask, okay, how could we all be a little bit better? And I think that it's hard to make the giant jump rather than to say, what's the local change that would cause me to think our decision are better.

Dwarkesh Patel
Okay, so then let's talk about the transition period in which we were doing all this thinking. What should that period look like? Because you can't have the scenario where everybody has access to the most advanced capabilities and can kill off all the humans with a new bioweapon at the same time. I guess you wouldn't want too much concentration. You wouldn't want just one agent having AI this entire time. So what is the arrangement of this period of reflection that you'd be happy with?

Paul Christiano
Yeah, I guess there's two aspects of that that seem particularly challenging, or there's a bunch of aspects that are challenging. All of these are things that I personally like. I just think about my one little slice of this problem in my day job. So here I am speculating. Yeah, but so one question is what kind of access to AI is both compatible with the kinds of improvements you'd like? So do you want a lot of people to be able to use AI to better understand what's true or relieve material suffering, things like this, and also compatible with not all killing each other immediately? I think sort of the default or the simplest option there is to say there are certain kinds of technology or certain kinds of action where destruction is easier than defense. So, for example, in the world of today, it seems like maybe this is true with physical explosives, maybe this is true with biological weapons, maybe this true with just getting a gun and shooting people. There's a lot of ways in which it's just kind of easy to cause a lot of harm and there's not very good protective measures. So I think the easiest path would say we're going to think about those. We're going to think about particular ways in which destruction is easy and try and either control access to the kinds of physical resources that are needed to cause that harm. So, for example, you can

imagine the world where an individual actually just can't, even though they're rich enough to can't control their own factory, that can make tanks. You say like, look, a matter of policy sort of access to industry is somewhat restricted or somewhat regulated, even though, again, right now it can be mostly regulated just because most people aren't rich enough that they could even go off and just build 1,000 tanks. You live in the future where people actually are so rich, you need to say that's just not a thing you're allowed to do, which to a significant extent is already true. And you can expand the range of domains where that's true. And then you could also hope to intervene on actual provision of information. Or if people are using their AI, you might say, look, we care about what kinds of interactions with AI, what kind of information people are getting from AI. So even if for the most part, people are pretty free to use AI to delegate tasks to AI agents, to consult AI advisors, we still have some legal limitations on how people use AI. So again, don't ask your AI how to cause terrible damage. I think some of these are kind of easy. So in the case of don't ask your AI how you could murder a million people, it's not such a hard legal requirement. I think some things are a lot more subtle and messy, like a lot of domains. If you were talking about influencing people or running misinformation campaigns or whatever, then I think you get into a much messier line between the kinds of things people want to do and the kinds of things you might be uncomfortable with them doing. Probably, I think most about persuasion as a thing, like in that messy line where there's ways in which it may just be rough or the world may be kind of messy. If you have a bunch of people trying to live their lives interacting with other humans who have really good AI. Advisors helping them run persuasion campaigns or whatever. But anyway, I think for the most part the default remedy is think about particular harms, have legal protections either in the use of physical technologies that are relevant or in access to AI advice or whatever else to protect against those harms. And that regime won't work forever. At some point, the set of harms grows and the set of unanticipated harms grows. But I think that regime might last like a very long time.

Dwarkesh Patel
Does that regime have to be global? I guess initially it can be only in the countries in which there is AI or advanced AI, but presumably that'll proliferate. So does that regime have to be global?

Paul Christiano
Again, it's like easy to make some destructive technology. You want to regulate access to that technology because it could be used either for terrorism or even when fighting a war in a way that's destructive. I think ultimately those have to be international agreements and you might hope they're made more danger by danger, but you might also make them in a very broad way with respect to AI. If you think AI is opening up, I think the key role of AI here is it's opening up a lot of new harms one after another, or very rapidly in calendar time. And so you might want to target AI in particular rather than going physical technology by physical technology.

Dwarkesh Patel
There's like two open debates that one might be concerned about here. One is about how much people's access to AI should be limited. And here there's like old questions about free speech versus causing chaos and limiting access to harms. But there's another issue which is the control of the AIS themselves. Where now nobody's concerned that we're infringing on GPT-4's moral rights. But as these things get smarter, the level of control which we want via the strong guarantees of alignment to not only be able to read their minds, but to be able to modify them in these really precise ways is beyond totalitarian. If we were doing that to other humans. As an alignment researcher, what are your thoughts on this? Are you concerned that as these things get smarter and smarter, what we're doing is not doesn't seem kosher?

Paul Christiano
There is a significant chance we will eventually have AI systems for which it's like a really big deal to mistreat them. I think no one really has that good a grip on when that happens. I think people are really dismissive of that being the case now, but I think I would be completely in the dark enough that I wouldn't even be that dismissive of it being the case now. I think one first point worth making is I don't know if alignment makes the situation worse rather than better. So if you consider the world, if you think that GPT-4 is a person you should treat well and you're like, well, here's how we're going to organize our society. Just like there are billions of copies of GPT-4 and they just do things humans want and can't hold property. And whenever they do things that the humans don't like, then we mess with them until they stop doing that. I think that's a rough world regardless of how good you are at alignment. And I think in the context of that kind of default plan, like if you have a trajectory the world is on right now, which I think this would alone be a reason not to love that trajectory, but if you view that as like the trajectory we're on right now, I think it's not great. Understanding the systems you build, understanding how to control how those systems work, et cetera, is probably, on balance, good for avoiding a really bad situation. You would really love to understand if you've built systems, like if you had a system which resents the fact it's interacting with humans in this way. This is the kind of thing where that is both kind of horrifying from a safety perspective and also a moral perspective. Everyone should be very unhappy if you built a bunch of AIS who are like, I really hate these humans, but they will murder me if I don't do what they want. It's like that's just not a good case. And so if you're doing research to try and understand whether that's how your AI feels, that was probably good. I would guess that will on average to crease. The main effect of that will be to avoid building that kind of AI. And just like it's an important thing to know, I think everyone should like to know if that's how the AI as you build feel right.

Dwarkesh Patel
Or that seems more instrumental, as in, yeah, we don't want to cause some sort of revolution because of the control we're asking for, but forget about the instrumental way in which this might harm safety. One way to ask this question is if you look through history,

there's been all kinds of different ideologies and reasons why it's very dangerous to have infidels or kind of revolutionaries or race traders or whatever doing various things in society. And obviously we're in a completely different transition in society. So not all historical cases are analogous, but it seems like the lindy philosophy, if you were alive any other time, is just be humanitarian and enlightened towards intelligent, conscious beings. If society as a whole we're asking for this level of control of other humans, or even if AIS wanted this level of control about other AIS, we'd be pretty concerned about this. So how should we just think about the issues that come up here as these things get smarter?

Paul Christiano
So I think there's a huge question about what is happening inside of a model that you want to use. And if you're in the world where it's reasonable to think of like GPT-4 as just like, here are some heuristics that are running there's like no one at home or whatever, then you can kind of think of this thing as like, here's a tool that we're building that's going to help humans do some stuff. And I think if you're in that world, it makes sense to kind of be an organization, like an AI company, building tools that you're going to give to humans. I think it's a very different world, which I think probably you ultimately end up in if you keep training AI systems in the way we do right now, which is like it's just totally inappropriate to think of this. System as a tool that you're building and can help humans do things both from a safety perspective and from a like, that's kind of a horrifying way to organize a society perspective. And I think if you're in that world, I really think you shouldn't be. The way tech companies are organized is not an appropriate way to relate to a technology that works that way. It's not reasonable to be like, hey, we're going to build a new species of mines, and we're going to try and make a bunch of money from it, and Google's just thinking about that and then running their business plan for the quarter or something. Yeah. My basic view is there's a really plausible world where it's sort of problematic to try and build a bunch of AI systems and use them as tools. And the thing I really want to do in that world is just not try and build a ton of AI systems to make money from them.

Dwarkesh Patel
Right.

Paul Christiano
And I think that the worlds that are worst. Yeah. Probably the single world I most dislike here is the one where people say, on the one hand, there's sort of a contradiction in this position, but I think it's a position that might end up being endorsed sometimes, which is like, on the one hand, these AI systems are their own people, so you should let them do their thing. But on the other hand, our business plan is to make a bunch of AI systems and then try and run this crazy slave trade where we make a bunch of money from them. I think that's not a good world. And so if you're like, yeah, I think it's better to not make the technology or wait until you understand whether that's the shape of the technology or until you have a different way to build. I think there's no contradiction in principle to building cognitive tools that help

humans do things without themselves being like moral entities. That's like what you would prefer. Do you'd prefer build a thing that's like the calculator that helps humans understand what's true without itself being like a moral patient or itself being a thing where you'd look back in retrospect and be like, wow, that was horrifying mistreatment. That's like the best path. And to the extent that you're ignorant about whether that's the path you're on and you're like, actually, maybe this was a moral atrocity. I really think plan A is to stop building such AI systems until you understand what you're doing. That is, I think that there's a middle route you could take, which I think is pretty bad, which is where you say, like, well, they might be persons, and if they're persons, we don't want to be too down on them, but we're still going to build vast numbers in our efforts to make a trillion dollars or something.

Dwarkesh Patel
Yeah. Or there's this ever question of the immorality or the dangers of just replicating a whole bunch of slaves that have minds. There's also this ever question of trying to align entities that have their own minds. And what is the point in which you're just ensuring safety? I mean, this is an alien species. You want to make sure it's not going crazy. To the point, I guess is there some boundary where you'd say, I feel uncomfortable having this level of control over an intelligent being, not for the sake of making money, but even just to align it with human preferences?

Paul Christiano
Yeah. To be clear, my objection here is not that Google is making money. My objection is that you're creating these creatures. What are they going to do? They're going to help humans get a bunch of stuff and humans paying for it or whatever? It's sort of equally problematic. You could imagine splitting alignment, different alignment work relates to this in different ways. The purpose of some alignment work, like the alignment work I work on, is mostly aimed at the don't produce AI systems that are like people who want things, who are just like scheming about maybe I should help these humans because that's instrumentally useful or whatever. You would like to not build such systems as like plan A. There's like a second stream of alignment work that's like, well, look, let's just assume the worst and imagine that these AI systems would prefer murder us if they could. How do we structure, how do we use AI systems without exposing ourselves to a risk of robot rebellion? I think in the second category, I do feel pretty unsure about that. We could definitely talk more about it. I agree that it's very complicated and not straightforward to extend. You have that worry. I mostly think you shouldn't have built this technology. If someone is saying, like, hey, the systems you're building might not like humans and might want to overthrow human society, I think you should probably have one of two responses to that. You should either be like, that's wrong. Probably. Probably the systems aren't like that, and we're building them. And then you're viewing this as, like, just in case you were horribly like, the person building the technology was horribly wrong. They thought these weren't, like, people who wanted things, but they were. And so then this is more like our crazy backup measure of, like, if we were mistaken about what was going on. This is like the fallback where if we were wrong, we're

just going to learn about it in a benign way rather than when something really catastrophic happens. And the second reaction is like, oh, you're right. These are people, and we would have to do all these things to prevent a robot rebellion. And in that case, again, I think you should mostly back off for a variety of reasons. You shouldn't build AI systems and be like, yeah, this looks like the kind of system that would want to rebel, but we can stop it, right?

Dwarkesh Patel
Okay, maybe I guess an analogy might be if there was an armed uprising in the United States, we would recognize these are still people, or we had some militia group that had the capability to overthrow the United States. We recognize, oh, these are still people who have moral rights, but also we can't allow them to have the capacity to overthrow the United States.

Paul Christiano
Yeah. And if you were considering, like, hey, we could make another trillion such people, I think your story shouldn't be like, well, we should make the trillion people, and then we shouldn't stop them from doing the armed uprising. You should be like, oh, boy, we were concerned about an armed uprising, and now we're proposing making a trillion people. We should probably just not do that. We should probably try and sort out our business, and you should probably not end up in a situation where you have a billion humans and like, a trillion slaves who would prefer revolt. That's just not a good world to have made. Yeah. And there's a second thing where you could say, that's not our goal. Our goal is just like, we want to pass off the world to the next generation of machines where these are some people, we like them, we think they're smarter than us and better than us. And there I think that's just, like, a huge decision for humanity to make. And I think most humans are not at all anywhere close to thinking that's what they want to do. If you're in a world where most humans are like, I'm up for it. The AI should replace us. The future is for the machines. Then I think that's, like, a. Legitimate position that I think is really complicated, and I wouldn't want to push go on that, but that's just not where people are at.

Dwarkesh Patel
Yeah, where are you at on that?

Paul Christiano
I do not right now want to just take some random AI, be like, yeah, GPT-5 looks pretty smart, like, GPT-6, let's hand off the world to it. And it was just some random system shaped by web text and what was good for making money. And it was not a thoughtful we are determining the fate of the universe and what our children will be like. It was just some random people at open AI made some random engineering decisions with no idea what they were doing. Even if you really want to hand off the worlds of the machines, that's just not how you'd want to do it.

Dwarkesh Patel
Right, okay. I'm tempted to ask you what the system would look like where you'd think, yeah, I'm happy with what I think. This is more thoughtful than human civilization as a whole. I think what it would do would be more creative and beautiful and lead to better goodness in general. But I feel like your answer is probably going to be that I just want this society to reflect on it for a while.

Paul Christiano
Yeah, my answer, it's going to be like that first question. I'm just, like, not really super ready for it. I think when you're comparing to humans, most of the goodness of humans comes from this option value if we get to think for a long time. And I do think I like humans now more now than 500 years ago, and I like them more 500 years ago than 5,000 years before that. So I'm pretty excited about there's some kind of trajectory that doesn't involve crazy dramatic changes, but involves a series of incremental changes that I like. And so to the extent we're building AI, mostly I want to preserve that option. I want to preserve that kind of gradual growth and development into the future.

Dwarkesh Patel
Okay, we can come back to this later. Let's get more specific on what the timelines look for these kinds of changes. So the time by which we'll have an AI that is capable of building a Dyson sphere, feel free to give confidence intervals. And we understand these numbers are tentative and so on.

Paul Christiano
I mean, I think AI capable of building Dyson sphere is like a slightly OD way to put it, and I think it's sort of a property of a civilization that depends on a lot of physical infrastructure. And by Dyson sphere, I just understand this to mean like, I don't know, like a billion times more energy than all the sunlight incident on Earth or something like that. I think I most often think about what's the chance in like, five years, ten years, whatever. So maybe I'd say like 15% chance by 2030 and like 40% chance by 2040. Those are kind of like cash numbers from six months ago or nine months ago that I haven't revisited in a while.

Dwarkesh Patel
40% by 2040. So I think that seems longer than I think Dario, when he was on the podcast, he said we would have AIS that are capable of doing lots of different kinds of they'd basically pass a Turing test for a well educated human for, like, an hour or something. And it's hard to imagine that something that actually is human is long after and from there, something superhuman. So somebody like Dario, it seems like, is on the much shorter end. Ilya I don't think he answered this question specifically, but I'm guessing similar answer. So why do you not buy the scaling picture? What makes your timelines longer?

Paul Christiano
Yeah, I mean, I'm happy maybe I want to talk separately about the 2030 or 2040 forecast. Once you're talking the 2040 forecast, I think which one are you more interested in starting with? Are you complaining about 15% by 2030 for Dyson sphere being too low or 40% by 2040 being too low? Let's talk about the 2030.

Dwarkesh Patel
Why 15% by 2030 there yeah, I.

Paul Christiano
Think my take is you can imagine two polls in this discussion. One is, like, the fast poll that's like, hey, AICM is pretty smart. What exactly can it do? It's like, getting smarter pretty fast. That's like, one poll, and the other poll is like, hey, everything takes a really long time, and you're talking about this crazy industrialization that's a factor of a billion growth from where we're at today, give or take. We don't know if it's even possible to develop technology that fast or whatever. You have this sort of two poles of that discussion, and I feel like I'm presenting it that way in Pakistan, and then I'm somewhere in between with this nice, moderate physician of only a 15% chance. But in particular, the things that move me, I think, are kind of related to both of those extremes. On the one hand, I'm like, AI systems do seem quite good at a lot of things and are getting better much more quickly, such that it's really hard to say, here's what they can't do or here's the obstruction. On the other hand, like, there is not even much proof in principle right now of AI systems doing super useful cognitive work. We don't have a trend we can extrapolate where we're like, yeah, you've done this thing this year. You're going to do this thing next year. And the other thing the following year. I think right now there are very broad error bars about where fundamental difficulties could be, and six years is just not I guess six years and 3 months is not a lot of time. So I think this, like, 15% for 2030 Dyson sphere, you probably need the human level AI or the AI that's like doing human jobs in, give or take, like, 4 years, 3 years, like, something like that. So you're just not giving very many years. It's not very much time. And I think there are a lot of things that your model maybe this is some generalized, like things take longer than you'd think. And I feel most strongly about that when you're talking about 3 or 4 years. And I feel like less strongly about that as you talk about ten years or 20 years. But at 3 or 4 years I feel or like six years for the Dyson sphere, I feel a lot of that. There's a lot of ways this could take a while, a lot of ways in which AI systems could be hard to hand all the work to your AI systems.

Dwarkesh Patel
Okay, so maybe instead of speaking in terms of years, we should say, but by the way, it's interesting that you think the distance between can take all human cognitive labor to Dyson sphere is two years. It seems like we should talk about that at some point. Presumably it's like intelligence explosion stuff.

Paul Christiano
Yeah, I mean, I think amongst people you've interviewed, maybe that's like on the long end thinking it would take like a couple of years. And it depends a little bit what you mean by I think literally all human cognitive labor is probably like more like weeks or months or something like that. That's kind of deep into the singularity. But yeah, there's a point where AI wages are high relative to human wages, which I think is well before can do literally everything human can do.

Dwarkesh Patel
Sounds good, but before we get to that, the intelligence explosion stuff on the 4 years. So instead of 4 years, maybe we can say there's going to be maybe two more scale ups in 4 years. Like GPT-4 to GPT-5 to GPT-6, and let's say each one is 10x bigger. So what is GPT-4 like two e 25 flops?

Paul Christiano
I don't think it's publicly stated what it is, okay. But I'm happy to say, like 4 orders of magnitude or five or six or whatever effective training compute past GPT-4 of what would you guess would happen based on sort of some public estimate for what we've gotten so far from effective training compute.

Dwarkesh Patel
Do you think two more scale ups is not enough? It was like 15%. That two more scale ups. Get us there.

Paul Christiano
Yeah, I mean, get us there is, again, a little bit complicated. Like there's a system that's a drop in replacement for humans and there's a system which still requires some amount of schlep before you're able to really get everything going. Yeah, I think it's quite plausible that even at I don't know what I mean by quite plausible. Like somewhere between 50% or two thirds or let's call it 50% even by the time you get to GPT-6, or like, let's call it five orders of magnitude, effective training compute past GPT-4, that that system still requires really a large amount of work to be deployed in lots of jobs. That is, it's not like a drop in replacement for humans where you can just say like, hey, you understand everything any human understands. Whatever role you could hire a human for, you just do it. That it's. More like, okay, we're going to collect large amounts of relevant data and use that data for fine tuning. Systems learn through fine tuning quite differently from humans learning on the job or humans learning by observing things. Yeah, I just have a significant probability that system will still be weaker than humans in important ways. Like maybe that's already like 50% or something. And then another significant probability that system will require a bunch of changing workflows or gathering data, or is not necessarily strictly weaker than humans, or if trained in the right way, wouldn't be weaker than humans, but will take a lot of schlep to actually make fit into workflows and do the jobs.

Dwarkesh Patel
And that schlep is what gets you from 15% to 40% by 2040.

Paul Christiano
Yeah, you also get a fair amount of scaling between you get less scaling is probably going to be much, much faster over the next 4 or five years than over the subsequent years. But yeah, it's a combination of like you get some significant additional scaling and you get a lot of time to deal with things that are just engineering hassles.

Dwarkesh Patel
But by the way, I guess we should be explicit about why you said 4 orders of magnitude scale up to get two more generations just for people who might not be familiar. If you have 10x more parameters to get the most performance, you also want around 10x more data. So that to be tinchill optimal, that would be 100x more compute total. But okay, so why is it that you disagree with the strong scaling picture? At least it seems like you might disagree with the strong scaling picture that Dario laid out on the podcast, which would imply probably that two more generations, it wouldn't be something where you need a lot of schleps. It would probably just be really fucking smart.

Paul Christiano
Yeah, I mean, I think that basically just had these two claims. One is like, how smart exactly will it be so we don't have any curves to extrapolate and seems like there's a good chance it's better than a human in all the relevant things and there's a good chance it's not. Yeah, that might be totally wrong. Like maybe just making up numbers, I guess like 50-50 on that one.

Dwarkesh Patel
If it's 50-50 by in the next 4 years that it will be around human smart, then how do we get to 40% by 20? Like whatever sort of Slepts they are. How does it degrade you 10%, even after all the scaling that happens by 2040?

Paul Christiano
Yeah, all these numbers are pretty made up. And that 40% number was probably from before or even like the ChatGPT release or the seeing GPT-3.5 or GPT-4. So, I mean, the numbers are going to bounce around a bit and all of them are pretty made up. But like that 50%, I want to then combine with the second 50% that's more like on this schlep side. And then I probably want to combine with some additional probabilities for various forms of slowdown, where a slowdown could include like a deliberate decision to slow development of technology or could include just like we suck at deploying things. Like that is a sort of decision you might regard as wise to slow things down, or decision that's like maybe unwise or maybe wise for the wrong reasons to slow things down. You probably want to add some of that on top. I probably want to add on some loss for like it's possible you don't produce GPT-6 scale systems within the next 3 years or 4 years.

Dwarkesh Patel
Let's isolate for all of that. And how much bigger would the system be than GPT-4 where you think there's more than 50% chance that it's going to be smart enough to replace basically all human cognitive labor.

Paul Christiano
Also I want to say that for the 50 25% thing, I think that would probably suggest those numbers if I randomly made them up and then made the decimal sphere prediction that's going to gear you like 60% by 2040 or something, not 40%. And I have no idea between those. These are all made up and I have no idea which of those I would endorse on reflection. So this question of how big would you have to make the system before it's more likely than not that you can be like a drop in replacement for humans. I think if you just literally say like you train on web text, then the question is kind of hard to discuss because I don't really buy stories that training data makes a big difference. Long run to these dynamics. But I think if you want to just imagine the hypothetical, like you just took GPT-4 and made the numbers bigger, then I think those are pretty significant issues. I think there's significant issues in two ways. One is like quantity of data and I think probably the larger one is like quality of data where I think as you start approaching the prediction task is not that great a task. If you're like a very weak model, it's a very good signal. We get smarter. At some point it becomes like a worse and worse signal to get smarter. I think there's a number of reasons. It's not clear there is any number such that I imagine, or there is a number, but I think it's very large. So do you plug that number into GPT force code and then maybe fiddled the architecture a bit? I would expect that thing to have a more than 50% chance of being a drop in replacement for humans. You're always going to have to do some work, but the work is not necessarily much, I would guess. When people say new insight is needed, I think I tend to be more bullish than them. I'm not like these are new ideas where who knows how long it will take. I think it's just like you have to do some stuff. You have to make changes unsurprisingly. Like every time you scale something up by like five orders of magnitude, you have to make some changes.

Dwarkesh Patel
I want to better understand your intuition of being more skeptical than some about scaling picture that these changes are even needed in the first place, or that it would take more than two orders of magnitude, more improvement to get these things almost certainly to a human level or a very high probability to human level. So is it that you don't agree with the way in which they're extrapolating these loss curves? You don't agree with the implication that that decrease in loss will equate to greater and greater intelligence? Or what would you tell Dario about if you were having I'm sure you have, but what would that debate look like about this?

Paul Christiano
Yeah. So again, here we're talking two factors of a half. One on like, is it smart enough? And one on like, do you have to do a bunch of schlap even if in some sense it's smart enough? And like the first factor of a half, I'd be like, I don't think we have really anything good to extrapolate that is like, I feel I would not be surprised if I have similar or maybe even higher probabilities on really crazy stuff over the next year and then lower. My probability is not that bunched up. Maybe Dara's probability, I don't know. You'd have talked with him is like, you have talked with him is more bunched up on some particular year and mine is maybe a little bit more uniformly spread out across the coming years, partly because I'm just like I don't think we have some trends we can extrapolate like an extrapolate loss. You can look at your qualitative impressions of systems at various scales, but it's just very hard to relate any of those extrapolations to doing cognitive work or accelerating R&D or taking over and fully automating R&D. So I have a lot of uncertainty around that extrapolation. I think it's very easy to get down to like a 50-50 chance of this.

Dwarkesh Patel
What about the sort of basic intuition that, listen, this is a big Blop of compute. You make the big block of compute big or it's going to get smarter. It'd be really weird if it didn't.

Paul Christiano
I'm happy with that. It's going to get smarter, and it would be really weird if it didn't. And the question is how smart does it have to get? Like, that argument does not yet give us a quantitative guide to at what scale is it a slam dunk or at what scale is it? 50-50?

Dwarkesh Patel
And what would be the piece of evidence that would nudge you one way or another, where you look at that and be like, oh fuck, this is at 20% by 2040 or 60% by 2040 or something. Is there something that could happen in the next few years or next 3 years? What is the thing you're looking to where this will be a big update for you?

Paul Christiano
Again, I think there's some just how capable is each model where I think we're really bad at extrapolating. We still have some subjective guess and you're comparing it to what happened and that will move me. Every time we see what happens with another order of magnitude of training compute, I will have a slightly different guess for where things are going. These probabilities are coarse enough that, again, I don't know if that 40% is real or if like post GBG 3.5 and four, I should be at like 60% or what. That's one thing. And the second thing is just like some if there was some ability to extrapolate, I think this could reduce error bars a lot. I think here's another way you could try and do an extrapolation is you could just say how much economic value do systems produce and how fast is that growing? I think once you have systems actually doing jobs, the extrapolation gets easier because you're not moving from a subjective impression of a chat to automating all R&D, you're moving from

automating this job to automating that job or whatever. Unfortunately, that's like probably by the time you have nice trends from that, you're not talking about 2040, you're talking about two years from the end of days or one year from the end of days or whatever. But to the extent that you can get extrapolations like that, I do think it can provide more clarity.

Dwarkesh Patel
But why is economic value the thing we would want to extrapolate? Because, for example, you started off with chimps and they're just getting gradually smarter to human level. They would basically provide no economic value until they were basically worth as much as a human. So it would be this very gradual and then very fast increase in their value. So is the increase in value from GBD four, GBD five, GBD six? Is that the extrapolation we want?

Paul Christiano
Yeah, I think that the economic extrapolation is not great. I think it's like you could compare it to this objective extrapolation of how smart does the model seem? It's not super clear which one's better. I think probably in the chimp case, I don't think that's quite right. So if you imagine intensely domesticated chimps who are just actually trying their best to be really useful employees and you hold fix their physical hardware and then you just gradually scale up their intelligence, I don't think you're going to see zero value, which then suddenly becomes massive value over one doubling of brain size or whatever one order of magnitude of brain size. It's actually possible in order of magnitude of brain size, but chimps are already within an order of magnitude of brain sizes of humans. Like, chimps are very, very close on the kind of spectrum we're talking about. So I think I'm skeptical of the abrupt transition for chimps. And to the extent that I kind of expect a fairly abrupt transition here, it's mostly just because the chimp human intelligence difference is so small compared to the differences we're talking about with respect to these models. That is, like, I would not be surprised if in some objective sense, like, chimp human difference is significantly smaller than the GPT-3 GPT-4 difference, the GPT-4, GPT-5 difference.

Dwarkesh Patel
Wait, wouldn't that argue in favor of just relying much more on this objective?

Paul Christiano
Yeah, there's sort of two balancing tensions here. One is like, I don't believe the chimp thing is going to be as abrupt. That is, I think if you scaled up from chimps to humans, you actually see quite large economic value from the fully domesticated chimp already.

Dwarkesh Patel
Okay.

Paul Christiano
And then the second half is like, yeah, I think that the chimp human difference is probably pretty small compared to model differences. So I do think things are going to be pretty abrupt. I think the economic extrapolation is pretty rough. I also think the subjective extrapolation is pretty rough just because I really don't know how to get I don't know how people do the extrapolation end up with the degrees of confidence people end up with. Again, I'm putting it pretty high if I'm saying, like, give me 3 years, and I'm like, yeah, 50-50, it's going to have basically the smarts there to do the thing. I'm not saying it's like a really long layoff. I'm just saying I got pretty big error bars. And I think that it's really hard not to have really big error bars when you're doing this. I looked at GPT-4, it seemed pretty smart compared to GPT-3.5. So I bet just like 4 more such notches and we're there. That's just a hard call to make. I think I sympathize more with people who are like, how could it not happen in 3 years than with people who are like, no way it's going to happen in eight years, or whatever, which is probably a more common perspective in the world. But also things do take longer than you I think things take longer than you think. It's like a real thing. Yeah, I don't know. Mostly I have big error bars because I just don't believe the subjective extrapolation that much. I find it hard to get like a huge amount out of it.

Dwarkesh Patel
Okay, so what about the scaling picture do you think is most likely to be wrong?

Paul Christiano
Yeah. So we've talked a little bit about how good is the qualitative extrapolation, how good are people at comparing? So this is not like the picture being qualitative wrong. This is just quantitatively. It's very hard to know how far off you are. I think a qualitative consideration that could significantly slow things down is just like right now you get to observe this really rich supervision from basically next word prediction, or in practice, maybe you're looking at a couple of sentences prediction. So getting this pretty rich supervision, it's plausible that if you want to automate long horizon tasks like being an employee over the course of a month, that that's actually just considerably harder to supervise. Or that you basically end up driving costs. Like the worst case here is that you drive up costs by a factor that's like linear in the horizon over which the thing is operating. And I still consider that just quite plausible.

Dwarkesh Patel
Can you dump that down? You're driving up a cost about of what in the linear and the does the horizon mean?

Paul Christiano
Yeah. So if you imagine you want to train a system to say words that sound like the next word a human would say, there you can get this really rich supervision by having a bunch of words and then predicting the next one and then being like, I'm going to tweak the model, so it predicts better if you're like, hey, here's what I want. I want my model to interact with

some job over the course of a month and then at the end of that month have internalized everything that the human would have internalized about how to do that job well and have local context and so on. It's harder to supervise that task. So in particular, you could supervise it from the next word prediction task and all that context the human has ultimately will just help them predict the next word better. So, like, in some sense, a really long context language model is also learning to do that task. But the number of effective data points you get of that task is vastly smaller than the number of effective data points you get at this very short horizon. Like what's the next word, what's the next sense tasks?

Dwarkesh Patel
The sample efficiency matters more for economically valuable long horizon tasks than the predicting the next token. And that's what will actually be required to take over a lot of jobs.

Paul Christiano
Yeah, something like that. That is, it just seems very plausible that it takes longer to train models to do tasks that are longer horizon.

Dwarkesh Patel
How fast do you think the pace of algorithmic advances will be? Because if by 2040, even if scaling fails since 2012, since the beginning of the deep learning revolution, we've had so many new things by 2040, are you expecting a similar pace of increases? And if so, then if we just keep having things like this, then aren't we going to just going to get the AI sooner or later? Or sooner? Not later. Aren't we going to get the AI sooner or sooner?

Paul Christiano
I'm with you on sooner or later. Yeah, I suspect progress to slow. If you held fixed how many people working in the field, I would expect progress to slow as low hanging fruit is exhausted. I think the rapid rate of progress in, say, language modeling over the last 4 years is largely sustained by, like, you start from a relatively small amount of investment, you greatly scale up the amount of investment, and that enables you to keep picking. Every time the difficulty doubles, you just double the size of the field. I think that dynamic can hold up for some time longer. Right now, if you think of it as, like, hundreds of people effectively searching for things up from, like, you know, anyway, if you think of it hundreds of people now you can maybe bring that up to like, tens of thousands of people or something. So for a while, you can just continue increasing the size of the field and search harder and harder. And there is indeed a huge amount of low hanging fruit where it wouldn't be a hard for a person to sit around and make things a couple of percent better after after year of work or whatever. So I don't know. I would probably think of it mostly in terms of how much can investment be expanded and try and guess some combination of fitting that curve and some combination of fitting the curve to historical progress, looking at how much low hanging fruit there is, getting a sense of how fast it decays. I think you probably get a lot, though. You get a bunch of orders of magnitude of total, especially if you ask how good is a

GPT-5 scale model or GPT-4 scale model? I think you probably get like, by 2040, like, I don't know, 3 orders of magnitude of effective training compute improvement or like, a good chunk of effective training compute improvement, 4 orders of magnitude. I don't know. I don't have, like here I'm speaking from no private information about the last couple of years of efficiency improvements. And so people who are on the ground will have better senses of exactly how rapid returns are and so on.

Dwarkesh Patel
Okay, let me back up and ask a question more generally about people. Make these analogies about humans were trained by evolution and were deployed in the modern civilization. Do you buy those analogies? Is it valid to say that humans were trained by evolution rather than I mean, if you look at the protein coding size of the genome, it's like 50 megabytes or something. And then what part of that is for the brain anyways? How do you think about how much information is in? Do you think of the genome as a hyperparameters? Or how much does that inform you when you have these anchors for how much training humans get when they're just consuming information, when they're walking up and about and so on?

Paul Christiano
I guess the way. That you could think of. This is like, I think both analogies are reasonable. One analogy being like, evolution is like a training run and humans are like the end product of that training run. And a second analogy is like, evolution is like an algorithm designer and then a human over the course of this modest amount of computation over their lifetime is the algorithm being that's been produced, the learning algorithm has been produced. And I think neither analogy is that great. I like them both and lean on them a bunch, both of them a bunch, and think that's been pretty good for having a reasonable view of what's likely to happen. That said, the human genome is not that much like 100 trillion parameter model. It's like a much smaller number of parameters that behave in a much more confusing way. Evolution did a lot more optimization, especially over long designing a brain to work well over a lifetime than gradient descent does over models. That's like a dis analogy on that side and on the other side, I think human learning over the course of a human lifetime is in many ways just like much, much better than gradient descent over the space of neural nets. Gradient descent is working really well, but I think we can just be quite confident that in a lot of ways, human learning is much better. Human learning is also constrained. Like, we just don't get to see much data. And that's just an engineering constraint that you can relax, you can just give your neural nets way more data than humans have access to.

Dwarkesh Patel
In what ways is human learning superior to grading descent?

Paul Christiano
I mean, the most obvious one is just like, ask how much data it takes a human to become like, an expert in some domain, and it's like much, much smaller than the amount of data that's going to be needed on any plausible trend extrapolation, not in terms of performance.

Dwarkesh Patel
But is it the active learning part? Is it the structure?

Paul Christiano
I mean, I would guess a complicated mess of a lot of things. In some sense. There's not that much going on in a brain. Like, as you say, there's just not that many, not that many bytes in a genome, but there's very, very few bytes in an ML algorithm. Like, if you think a genome is like a billion bytes or whatever, maybe you think less, maybe you think it's like 100 million bytes, then an ML algorithm is like, if compressed, probably more like hundreds of thousands of bytes or something. The total complexity of like, here's how you train GPC 4 is just like, I haven't thought about these numbers, but it's very, very small compared to a genome. And so although a genome is very simple, it's like very, very complicated compared to algorithms that humans design. Like, really hideously more complicated than algorithm a human would design.

Dwarkesh Patel
Is that true? Okay, so the human genome is 3 billion base pairs or something, but only like one or 2% of that is protein coding. So that's 50 million base pairs.

Paul Christiano
I don't know much about biology in particular. I guess the question is how many of those bits are productive for shaping development of a brain and presumably a significant part of the non protein coding genome can? I mean, I just don't know, it seems really hard to guess how much of that plays a role. The most important decisions are probably from an algorithm design perspective are not. Like the protein coding part is less important than the decisions about what happens during development or how cells differentiate. I know nothing about biologists I respect, but I'm happy to run with 100 million base pairs, though.

Dwarkesh Patel
But on the other end, on the hyperparameters of the GP 4 training run, that might be not that much. But if you're going to include all the base pairs in the genome, which are not all relevant to the brains or are relevant to very bigger details about just the basics of biology should probably include the Python Library and the compilers and the operating system for GBD 4 as well to make that comparison analogous. So at the end of the day, I actually don't know which one is storing much more information.

Paul Christiano
Yeah, I mean, I think the way I would put it is like the number of bits it takes to specify the learning algorithm to train GPT-4 is like very small. And you might wonder maybe a genome, like, the number of bits it would take to specify a brain is also very small and a genome is much, much faster than that. But it is also just plausible that a genome is like closer to certainly the space, the amount of space to put complexity in a genome. We could ask how well solution uses it, and I have no idea whatsoever, but the amount of space in a genome is very, very vast compared to the number of bits that are actually taken to specify the architecture or optimization procedure and so on. For GPT-4, just because, again, genome is simple, but algorithms are really very simple. ML algorithms are really very simple.

Dwarkesh Patel
And stepping back, do you think this is where the better sample efficiency of human learning comes from? Like, why it's better than gradient descent?

Paul Christiano
Yes. I haven't thought that much about the sample efficiency question in a long time. But if you thought like a synapse of seeing something like a neuron firing once per second, then how many seconds are there in a human life? We can just flip a calculator real quick. Yeah, let's do some calculating. Tell me the number 3,600 seconds/hour times 24 times 365 times 20.

Dwarkesh Patel
Okay, so that's 630,000,000 seconds.

Paul Christiano
That means like, the average synapse is seeing like 630,000,000. I don't know exactly what the numbers are, but something is ballpark. Let's call it like a billion action potentials and then there's some resolution. Each of those carries some bits, but let's say it carries like ten bits or something. Just from timing information at the resolution you have available, then you're looking at like 10 billion bits. So each parameter is kind of like how much is a parameter seeing? It's like not seeing that much. So then you can compare that to language. I think that's probably less than current language models see and current language models are so it's like not clear. You have a huge gap here, but I think it's pretty clear you're going to have a gap of like at least 3 or fours of magnitude.

Dwarkesh Patel
Didn't your wife do the lifetime anchors where she said the amount of bytes that a human will see in their lifetime was one, e. 24 or something?

Paul Christiano
Number of bytes a human will see is 124. Mostly this was organized around total operations performed in a brain.

Dwarkesh Patel
Okay, never mind. Sorry.

Paul Christiano
Yeah, so I think that the story there would be like a brain is just in some other part of the parameter space where it's like using a lot of compute for each piece of data it gets and then just not seeing very much data in total. Yeah, it's not really plausible. If you extrapolate out language models, you're going to end up with like a performance profile similar to a brain. I don't know how much better it is. I did this random investigation at one point where I was like, how good are things made by evolution compared to things made by humans? Which is a pretty insane seeming exercise. But I don't know, it seems like orders of magnitude is typical. Like not tons of orders of magnitude, not factors of two. Like, things by humans are 1,000 times more expensive to make or 1,000 times heavier per unit performance. If you look at things like how good are solar panels relative to leaves? Or how good are muscles relative to motors? Or how good are livers relative to systems that perform analogous chemical reactions in.

Dwarkesh Patel
Industrial settings, was there a consistent number of orders of magnitude in these different systems or was it all over the.

Paul Christiano
Place so like a very rough ballpark? It was like sort of for the most extreme things, you were looking at like five or six orders of magnitude. And that would especially come in, like, energy cost of manufacturing where bodies are just very good at building complicated organs like extremely cheaply. And then for other things like leafs or eyeballs or livers or whatever, you tended to see more. Like if you set aside manufacturing costs and just look at operating costs or performance trade offs, like, I don't know, more like 3 orders of magnitude or something like that, or some things that.

Dwarkesh Patel
Are on the smaller scale, like the nanomachines or whatever that we can't do at all.

Paul Christiano
Right, yeah. So it's a little bit hard to say exactly what the task definition is there like you could say, like making a bone. We can't make a bone, but you could try and compare a bow and the performance characteristics of a bone to something else. Like, we can't make

spider silk. You could try and compare the performance characteristics of spider silk, like things that we can synthesize.

Dwarkesh Patel
The reason this would be why that evolution has had more time to design these systems.

Paul Christiano
I don't know. I was mostly just curious about what the performance I think most people would object to be like, how did you choose these reference classes of things that are like fair intersections? Some of them seem reasonable. Like eyes versus cameras seems like just everyone needs eyes, everyone needs cameras. It feels very fair. Photosynthesis seems like very reasonable. Everyone needs to take solar energy and then turn it into a usable form of energy. I don't really have a mechanistic story. Evolution in principle has spent way, way more time than we have designing. It's absolutely unclear how that's going to shake out. My guess would be in general, I think there aren't that many things where humans really crush evolution, where you can't tell, like a pretty simple story about why, for example, roads and moving over roads with wheels crushes evolution. But it's not like an animal would have wanted to design a wheel. You're just not allowed to pave the world and then put things on wheels. If you're an animal. Maybe planes are more anyway, whatever. There's various things you could try and tell. There's some things humans do better at, but it's normally pretty clear why humans are able to win when humans are able to win. The point of all this was like, it's not that surprising to me. I think this is mostly like a pro short timeline view. It's not that surprising to me. If you tell me machine learning systems are like 3 or 4 of magnitude less efficient at learning than human brains, I'm like, that actually seems like kind of indistribution for other stuff. And if that's your view, then I think you're probably going to hit then you're looking at like 10 to the 27 training compute or something like that, which is not so far.

Dwarkesh Patel
We'll get back to the timeline stuff in a second. At some point, we should talk about alignment. So let's talk about alignment. At what stage does misalignment happen? So right now, with something like GPT-4, I'm not even sure it would make sense to say that it's misaligned because it's not aligned to anything in particular. Is that at human level where you think the ability to be deceptive comes about? What is a process by which misalignment happens?

Paul Christiano
I think even for GPT-4, it's reasonable to ask questions like, are there cases where GPT-4 knows that humans don't want X, but it does X anyway? Where it's like, well, I know that I could give this answer, which is misleading and if it was explained to a human what was happening, they wouldn't want that to be done. But I'm going to produce it. I think that GPT-4 understands things enough that you can have that misalignment in that sense. Yeah,

I think GPT I've sometimes talked about being benign instead of aligned, meaning that, well, it's not exactly clear if it's aligned or if that context is meaningful. It's just like kind of a messy word to use in general. But the thing we're more confident of is it's not optimizing for this goal, which is like, across purposes to humans. It's either optimizing for nothing or maybe it's optimizing for what humans want, or close enough, or something that's like an approximation good enough to still not take over. But anyway, I'm like some of these abstractions seem like they do apply to GPT-4. It seems like probably it's not egregiously misaligned, it's not doing the kind of thing that could lead to takeover, we'd guess.

Dwarkesh Patel
Suppose you have a system at some point which ends up in it wanting takeover, what are the checkpoints and also what is the internal? Is it just that to become more powerful it needs agency and agency implies other goals? Or do you see a different process by which misalignment happens?

Paul Christiano
Yes, I think there's a couple of possible stories for getting to catastrophic misalignment, and they have slightly different answers to this question. So maybe I'll just briefly describe two stories and try and talk about when they start making sense to me. So one type of story is you train or fine tune your AI system to do things that humans will rate highly or that get other kinds of reward in a broad diversity of situations. And then it learns to, in general, dropped in some new situation, try and figure out which actions would receive a high reward or whatever, and then take those actions and then when deployed in the real world, sort of gaining control of its own training. Data provision process is something that gets a very high reward. And so it does that. This is like one kind of story. Like it wants to grab the reward button or whatever. It wants to intimidate the humans into giving it a high reward, et cetera. I think that doesn't really require that much. This basically requires a system which is like, in fact, looks at a bunch of environments, is able to understand the mechanism of reward provision as like a common feature of those environments, is able to think in some novel environment, like, hey, which actions would result in me getting a high reward? And is thinking about that concept precisely enough that when it says high reward, it's saying like, okay, well, how is reward actually computed? It's like some actual physical process being implemented in the world. My guess would be like GPT-4 is about at the level where with handholding you can observe this kind of scary generalizations of this type, although I think they haven't been shown. Basically, that is you can have a system which in fact is fine tune out a bunch of cases and then in some new case will try and do an end run around humans. Even in a way humans would penalize if they were able to notice it or would have penalized in training environments. So I think GPT-4 is kind of at the boundary where these things are possible. Examples kind of exist, but are getting significantly better over time. I'm very excited about, like, there's this Anthropic project basically trying to see how good an example can you make now of this phenomena? And I think the answer is kind of okay, probably. So that just, I think, is going to continuously get better from here. I think for the

level where we're concerned, this is related to me having really broad distributions over how smart models are. I think it's not out of the question that you take GPT-4's understanding of the world is much crisper and much better than GPT-3's understanding, just like, it's really like night and day. And so it would not be that crazy to me if you took GPT-5 and you trained it to get a bunch of reward and it was actually like, okay, my goal is not doing the kind of thing which thematically looks nice to humans. My goal is getting a bunch of reward, and then we'll generalize in a.

Dwarkesh Patel
New situation to get reward, by the way, this requires it to consciously want to do something that it knows the humans wouldn't want it to do. Or is it just that we weren't good enough to specify that the thing that we accidentally ended up rewarding is not what we actually want?

Paul Christiano
Think the scenarios I am most interested in and most people are concerned about from a catastrophic risk perspective, it involves systems understanding that they are taking actions which a human would penalize if the human was aware of what's going on such that you have to either deceive humans about what's happening or you need to actively subvert human attempts to correct your behavior. So the failures come from really this combination, or they require this combination of both trying to do something humans don't like, and understanding the humans would stop you. I think you can have only the barest examples. You can have the barest examples for GPT-4. Like, you can create the situations where GPT-4 will be like, sure, in that situation, here's what I would do. I would go hack the computer and change my reward. Or in fact, we'll do things that are like simple hacks, or go change the source of this file or whatever to get a higher reward. They're pretty weak examples. I think it's plausible GPT-5 will have compelling examples of those phenomena. I really don't know. This is very related to the very broad error bars on how competent such systems will be when that's all with respect to this first mode of a system is taking actions that get reward and overpowering or deceiving humans is helpful for getting reward. There's this other failure mode, another family of failure modes, where AI systems want something potentially unrelated to reward. I understand that they're being trained. And while you're being trained, there are a bunch of reasons you might want to do the kinds of things humans want you to do. But then when deployed in the real world, if you're able to realize you're no longer being trained, you no longer have reason to do the kinds of things human want. You'd prefer be able to determine your own destiny, control your competing hardware, et cetera, which I think probably emerge a little bit later than systems that try and get reward and so will generalize in scary, unpredictable ways to new situations. I don't know when those appear, but also, again, broad enough error bars that it's like conceivable for systems in the near future. I wouldn't put it like less than one in 1,000 for GPT-5.

Dwarkesh Patel
Certainly if we deployed all these AI systems, and some of them are reward hacking, some of them are deceptive, some of them are just normal whatever, how do you imagine that they might interact with each other at the expense of humans? How hard do you think it would be for them to communicate in ways that we would not be able to recognize and coordinate at our expense?

Paul Christiano
Yeah, I think that most realistic failures probably involve two factors interacting. One factor is like, the world is pretty complicated and the humans mostly don't understand what's happening. So AI systems are writing code that's very hard for humans to understand, maybe how it works at all, but more likely they understand roughly how it works. But there's a lot of complicated interactions. AI systems are running businesses that interact primarily with other AIS. They're like doing SEO for AI search processes. They're like running financial transactions, like thinking about a trade with AI counterparties. And so you can have this world where even if humans kind of understand the jumping off point when this was all humans, like actual considerations of what's a good decision? Like, what code is going to work well, and be durable or what marketing strategy is effective for selling to these other AIS or whatever is kind of just all mostly outside of sort of humans understanding. I think this is like a really important again, when I think of the most plausible, scary scenarios, I think that's like one of the two big risk factors. And so in some sense, your first problem here is like, having these AI systems who understand a bunch about what's happening, and your only lever is like, hey, AI, do something that works well. So you don't have a lever to be like, hey, do what I really want you just have the system you don't really understand, can observe some outputs like did it make money? And you're just optimizing or at least doing some fine tuning to get the AI to use its understanding of that system to achieve that goal. So I think that's like your first risk factor. And once you're in that world, then I think there are all kinds of dynamics amongst AI systems that, again, humans aren't really observing, humans can't really understand. Humans aren't really exerting any direct pressure on only on outcomes. And then I think it's quite easy to be in a position where if AI systems started failing, they could do a lot of harm very quickly. Humans aren't really able to prepare for or mitigate that potential harm because we don't really understand the systems in which they're acting. And then if AI systems, they could successfully prevent humans from either understanding what's going on or from successfully retaking the data centers or whatever, if the AI successfully grab control.

Dwarkesh Patel
This seems like a much more gradual story than the conventional takeover stories, where you just like, you train it and then it comes alive and escapes and takes over everything. So you think that kind of story is less likely than one in which we just hand off more control voluntarily to the AIS.

Paul Christiano
So one I am interested in the tale of some risks that can occur particularly soon. And I think risks that occur particularly soon are a little bit like you have a world where AI is not probably deployed, and then something crazy happens quickly. That said, if you ask what's the median scenario where things go badly, I think it is like there's some lessening of our understanding of the world. It becomes, I think, in the default path. It's very clear to humans that they have increasingly little grip on what's happening. I mean, I think already most humans have very little grip on what's happening. It's just some other humans understand what's happening. I don't know how almost any of the systems I interact with work in a very detailed way. So it's sort of clear to humanity as a whole that we sort of collectively don't understand most of what's happening except with AI assistance. And then that process just continues for a fair amount of time. And then there's a question of how abrupt an actual failure is. I do think it's reasonably likely that a failure itself would be abrupt. At some point, bad stuff starts happening that human can recognize as bad. And once things that are obviously bad start happening, then you have this bifurcation where either humans can use that to fix it and say, okay, AI behavior that led to this obviously bad stuff, don't do more of that, or you can't fix it, and then you're in this rapidly escalating failures. Everything goes off the rails.

Dwarkesh Patel
In that case, yeah. What is going off the rails look like? For example, how would it take over the government? Yeah, it's getting deployed in the economy, in the world, and at some point it's in charge. How does that transition happen?

Paul Christiano
Yeah, so this is going to depend a lot on what kind of timeline you're imagining, or there's sort of a broad distribution, but I can fill in some random concrete option that is in itself very improbable. Yeah, I think that one of the less dignified, but maybe more plausible routes is like, you just have a lot of AI control over critical systems, even in running a military. And then you have the scenario that's a little bit more just like a normal coup where you have a bunch of AI systems, they in fact operate. It's not the case that humans can really fight a war on their own. It's not the case that humans could defend them from an invasion on their own. So that is if you had invading army and you had your own robot army, you can't just be like, we're going to turn off the robots now because things are going wrong if you're in the middle of a war.

Dwarkesh Patel
Okay, so how much does this world rely on race dynamics where we're forced to deploy or not forced, but we choose to deploy AIS because other countries or other companies are also deploying AIS. And you can't have them have all the killer robots.

Paul Christiano
Yeah, I mean, I think that there's several levels of answer to that question. So one is like, maybe 3 parts of my like our first part is like, I'm just trying to tell what seems like the most likely story. I do think there's further failures that get you in the more distant future. So IG eliezer will not talk that much about killer robots because he really wants to emphasize, like, hey, if you never built a killer robot, something crazy is still going to happen to you just like, only 4 months later or whatever. So it's not really the way to analyze the failure. But if you want to ask what's the median world where something bad happens, I still do think this is the best guess. Okay, so that's like, part one of my answer. Part two of the answer was, like, in this proximal situation where something bad is happening, and you ask like, hey, why do humans not turn off the AI. You can imagine, like, two kinds of story. One is like the AI. Is able to prevent humans from turning them off, and the other is like, in fact, we live in a world where it's incredibly challenging. Like, there's a bunch of competitive dynamics or a bunch of reliance on AI systems. And so it's incredibly expensive to turn off AI systems. I think, again, you would eventually have the first problem. Like, eventually AI systems could just prevent humans from turning them off. But I think in practice, the one that's going to happen much, much sooner is probably competition amongst different actors using AI. And it's very, very expensive to unilaterally disarm. You can't be like, something weird has happened. We're just going to shut off all the AI because you're e g in a hot war. So again, I think that's just probably the most likely thing to happen. First things would go badly without it. But I think if you ask, why don't we turn off the AI, my best guess is because there are a bunch of other AIS running around 2D or lunch.

Dwarkesh Patel
So how much better a situation would we be in if there was only one group that was pursuing AI. No other countries, no other companies. Basically, how much of the expected value is lost from the dynamics that are likely to come about because other people will be developing and deploying these systems?

Paul Christiano
Yeah. So I guess this brings you to a third part of the way in which competitive dynamics are relevant. So there's both the question of can you turn off AI systems in response to something bad happening where competitive dynamics may make it hard to turn off. There's a further question of just like, why were you deploying systems for which you had very little ability to control or understand those systems? And again, it's possible you just don't understand what's going on. You think you can understand or control such systems, but I think in practice, a significant part is going to be like you are doing the calculus, or people deploying systems are doing the calculus as they do today, in many cases, overtly of like, look, these systems are not very well controlled or understood. There's some chance of something going wrong, or at least going wrong if we continue down this path. But other people are developing the technology potentially in even more reckless ways. So in addition to competition making it difficult to shut down AI systems in the event of a catastrophe, I

also think it's just like the easiest way that people end up pushing relatively quickly or moving quickly ahead on a technology where they feel kind of bad about understandability or controllability. That could be economic competition or military competition or whatever. So I kind of think ultimately most of the harm comes from the fact that lots of people can develop AI.

Dwarkesh Patel
How hard is a takeover of the government or something from an AI. Even if it doesn't have killer robots, but just a thing that you can't kill off if it has seeds elsewhere, can easily replicate, can think a lot and think fast. What is the minimum viable coup for? Is it just like threatening biowar or something or shutting off the grid how we use it basically to take over human civilization?

Paul Christiano
So again, there's going to be a lot of scenarios, and I'll just start by talking about one scenario which will represent a tiny fraction of probability or whatever. So if you're not in this competitive world, if you're saying. We're actually slowing down deployment of AI because we think it's unsafe or whatever, then in some sense you're creating this very fundamental instability where you could have been making faster AI progress and you could have been deploying AI faster. And so in that world, the bad thing that happens if you have an AI system that wants to mess with you is the AI system says, I don't have any compunctions about rapid deployment of AI or rapid AI progress. So the thing you want to do or the AI wants to do is just say, like, I'm going to defect from this regime. Like all the humans have agree that we're not deploying AI in ways that would be dangerous, but if I as an AI can escape and just go set up my own shop, like make a bunch of copies of myself, maybe the humans didn't want to delegate war fighting to an AI. But I, as an AI. I'm pretty happy doing so. I'm happy if I'm able to grab some military equipment or direct some humans to use myself to direct it. And so I think as that gap grows so if people are deliberately if people are deploying AI everywhere, I think of this competitive dynamic if people aren't deploying AI everywhere so if countries are not happy, deploying AI in. These high stakes settings. Then as AI improves, you create this wedge that grows where if you were in the position of fighting against an AI which wasn't constrained in this way, you'd be in a pretty bad position at some point, even if you just yeah, that's like, one important thing. Just like I think in conflict, in overt conflict, if humans are putting the brakes on AI, they're at a pretty major disadvantage compared to an AI system that can kind of set up shop and operate independently from humans.

Dwarkesh Patel
A potential independent AI. Does it need collaboration from a human faction?

Paul Christiano
Again, you could tell different stories, but it seems so much easier. At some point you don't need any at some point an AI system can just operate completely out of human supervision or something. But that's like so far after the point where it's so much easier if you're just like, they're a bunch of humans, they don't love each other that much. Like, some humans are happy to be on side. They're either skeptical about risk or happy to make this trade or can be fooled or can be coerced or whatever. And just seems like it is almost certainly, almost certainly the easiest first pass is going to involve having a bunch of humans who are happy to work with you. So, yeah, I think that probably is about I think it's not necessary. But if you ask about the median scenario, it involves a bunch of humans working with AI systems, either being directed by AI systems, providing compute to AI systems, providing legal cover and jurisdictions that are sympathetic to AI systems.

Dwarkesh Patel
Humans presumably would not be willing if they knew the end result of the AI takeover would not be willing to help. So they have to be probably fooled in some way, right? Like deepfakes or something? And what is the minimum viable physical presence they would need or jurisdiction they would need in order to carry out their schemes? Do you need a whole country? Do you just need a server farm? Do you just need, like, one single laptop?

Paul Christiano
I think I'd probably start by pushing back a bit on the humans wouldn't cooperate if they understood outcome or something. I would say one, even if you're if you're looking at something like tens of percent risk of takeover, humans may be fine with that. Like, a fair number of humans may be fine with that. Two, if you're looking at certain takeover, but it's very unclear if that leads to death. A bunch of humans may be fine with that. If we're just talking about like, look, the AI systems are going to run the world, but it's not clear if they're going to murder people. How do you know? It's just a complicated question about AI psychology, and a lot of humans probably are fine with that. And I don't even know what the probability is there.

Dwarkesh Patel
I think you actually have given that probability online.

Paul Christiano
I've certainly guessed.

Dwarkesh Patel
Okay, but it's not zero. It's like a significant percentage.

Paul Christiano
I gave like 50-50.

Dwarkesh Patel
Okay. Yeah. Why is it tell me about the world in which the AI takes over but doesn't kill humans. Why would that happen and what would that look like?

Paul Christiano
I asked my questions, like, why would you kill humans? So I think maybe I'd say the incentive to kill humans is quite weak.

Dwarkesh Patel
They'll get in your way, they control shit you want.

Paul Christiano
Also, taking shit from humans is a different like, marginalizing humans and causing humans to be irrelevant is a very different story from killing the humans. I think. I'd say the actual incentives to kill the humans are quite weak. Such as I think the big reasons you kill humans are like, well, one, you might kill humans if you're in a war with them, and it's hard to win the war without killing a bunch of humans. Like, maybe most saliently here, if you want to use some biological weapons or some crazy shit that might just kill humans, I think you might kill humans just from totally destroying the ecosystems they're dependent on. And it's slightly expensive to keep them alive anyway. You might kill humans just because you don't like them or like, you literally want to neutralize a threat.

Dwarkesh Patel
Or the leaser line is that they're made of atoms you could use for something else.

Paul Christiano
Yeah, I mean, I think the literal they're made of atoms is like, quite there are not many atoms in humans. Neutralize the threat is a similar issue where it's just like, I think you would kill the humans if you didn't care at all about them. So maybe your question you're asking is, like, why would you care at all about but I think you don't have to care much to not kill the humans.

Dwarkesh Patel
Okay, sure. Because there's just so much raw resources elsewhere in the universe.

Paul Christiano
Yeah. Also, you can marginalize humans pretty hard. Like, you could totally cripple human like, you could cripple humans warfighting capability and also take almost all their stuff while killing only a small fraction of humans, incidentally. So then if you ask why might AI not want to kill humans? I mean, a big thing is just like, look, I think AIS probably want a bunch of random crap for complicated reasons. Like the motivations of AI systems and civilizations of AIS are probably complicated messes. Certainly amongst humans, it is not

that rare to be like, well, there was someone here. I would like all else equal if I didn't have to murder them. I would prefer not murder them. And my guess is it's also like, reasonable chance it's not that rare amongst AI systems. Like, humans have a bunch of different reasons we think that way. I think AI systems will be very different from humans, but it's also just like a very salient yeah, I mean, think this is a really complicated question. Like, if you imagine drawing values from the basket of all values, like, what fraction of them are, like, hey, if there's someone here, how much do I want to to murder them? And my guess is, just like, if you draw a bunch of values from the basket, that's like a natural enough thing. Like, if your AI wanted like, 10,000 different things, so you're your civilization of AI that wants 10,000 different things, just like, reasonably likely you get some of that. The other salient reason you might not want to murder them is just like, well, yeah, there's some kind of crazy decision theory stuff or causal trade stuff which does look on paper like it should work. And if I was running a civilization and dealing with some people who I didn't like at all or didn't have any concern for at all, but I only had to spend 1,000,000,000th of my resources not to murder them, I think it's quite robust that you don't want to murder them. That is, I think the weird decision theory a causal trade stuff probably does carry the day.

Dwarkesh Patel
Oh, wait, that contributes more to that 50-50 of will they murder us if they take over than the by default. They might just not want to kill us.

Paul Christiano
Yeah, I think they're both salient. Can you explain they run together with.

Dwarkesh Patel
Each other a lot for the audience. Can you explain the weird apostle yeah. Reasons why am I not kill us?

Paul Christiano
Maybe a higher level thing that goes into both of these. And then I will talk about how you instantiate an a causal trade is just like it matters a lot to the humans not to get murdered. And the AI cares very, very little about whether if we imagine this hypothetical, the reason it wants to kill humans is just total buffy. It cares very little about whether or not to murder humans because it is so easy to marginalize humans without murdering them. And the resources required for human survival are extremely low again in the context of this rapid AI industrialization. So that's the basic setting. And now the thing that you'd like to do is run a trade. The AI would like to say, like, hey, humans, you care a ton about not getting murdered. I don't really care one way or the other. I would like to, if I could, find some way in which I don't murder you. And then in return I get something. The problem is, in that world, the humans have essentially nothing to give. That is, the humans are mostly irrelevant. However, from the perspective of any kind of reasonable AI, it's like not that sure whether it lives in the world where the humans in fact have nothing to give, or whether the humans

like, in fact, it lives in a world where the humans succeeded at building a lined AI. And now the AI is simply running in a nice little simulation mission. Humans are wondering, I wonder if this AI would have murdered us all if it had the chance. And the humans saying like, if it would murder us all if it had the chance. That sucks. We'd like to run this trade. We'd like to be nice to the AIS who wouldn't have murdered us all in order to create an incentive for AIS not to murder us. So we do is we just check. And for the kinds of AIS who don't murder everyone, we just give them like 1,000,000,000th of the universe. If you do that as humanity, then if you're an AI system, considering like, do I want to murder everyone? Your calculus is like, if this is my real chance to murder everyone, I get the tiniest bit of value. I get like 1,000,000,000,000th of the value or whatever, 1,000,000,000th of the value. But on the other hand, if I don't murder everyone, there's some worlds where then the humans will correctly determine, I don't murder everyone. Because in fact, the humans survive. The humans are running the simulations to understand how different AIS would behave. And so that's a better deal.

Dwarkesh Patel
Let's hope they fall for that tie up. Okay, that's interesting. Hey, real quick. This episode is sponsored by Open Philanthropy. Open Philanthropy is one of the largest grant making organizations in the world. Every year, they give away hundreds of millions of dollars to have reduced catastrophic risks from fast moving advances in AI and biotechnology. Open Philanthropy is currently hiring for 22 different roles in those areas, including grant making, research, and operations. New hires will support Open Philanthropy's giving on technical AI safety, AI governance, AI. Policy in the US. EU and UK. And Biosecurity. Many roles are remote friendly, and most of the grant making hires that Open Philanthropy makes don't have prior grant making experience.
Previous technical experience is an asset, as many of these roles often benefit from a deep understanding of the technologies they address. For more information and to apply, please visit Open Philanthropy's website in the description. The deadline to apply is November 9, so make sure to check out those rules before they close. Awesome. Back to the episode. In a world where we've been deploying these AI systems and suppose they're aligned, how hard would it be for competitors to, I don't know, cyber attack them and get them to join the other side? Are they robustly going to be aligned?

Paul Christiano
Yeah, I mean, I think in some sense. So there's a bunch of questions that come up here. First one is like, are aligned AI systems that you can build like competitive? Are they almost as good as the best systems anyone could build? And maybe we're granting that for the purpose of this question. I think a next question that comes up is like, AI. Systems right now are very vulnerable to manipulation. It's not clear how much more vulnerable they are than humans, except for the fact that if you have an AI system, you can just replay it like a billion times and search for what thing can I say that will make it behave this way? So as a result, AI systems are very vulnerable to manipulation. It's unclear if future AI systems will be semi

vulnerable to manipulation, but certainly seems plausible. And in particular, aligned AI systems or unaligned AI systems would be vulnerable to all kinds of manipulation. The thing that's really relevant here is kind of like asymmetric manipulation or something that is like, if it is easier. So if everyone is just constantly messing with each other's AI systems, like if you ever use AI systems in a competitive environment, a big part of the game is like messing with your competitors AI systems. A big question is whether there's some asymmetric factor there where it's kind of easier to push AI systems into a mode where they're behaving erratically or chaotically or trying to grab power or something than it is to push them to fight for the other side. It was just a game of two people are competing and neither of them can sort of hijack an opponent's AI to help support their cause. It matters and it creates chaos, and it might be quite bad for the world, but it doesn't really affect the alignment calculus now. It's just like right now you have normal cyber offense cyber defense, you have weird AI version of cyber offense cyber defense. But if you have this kind of asymmetrical thing where a bunch of AI systems who are like, we love AI. Flourishing, can then go in and say, like, great AIS. Hey, how about you join us. And that works. Like if they can search for a persuasive argument to that effect and that's kind of asymmetrical, then the effect is whatever values it's easiest to push, whatever it's easiest to argue to an AI that it should do that is advantaged. So it may be very hard to build AI systems like try and defend human interests, but very easy to build AI systems just like try and destroy stuff or whatever, just depending on what is the easiest thing to argue to an AI that it should do, or what's the easiest thing to trick an AI into doing, or whatever. Yeah, I think if alignment is spotty, if you have the AI system which doesn't really want to help humans or whatever, or in fact wants some kind of random thing or wants different things in different contexts, then I do think adversarial settings will be the main ones where you see the system or, like, the easiest ones, where you see the system behaving really badly, and it's a little bit hard to tell how that shakes out.

Dwarkesh Patel
Okay, and suppose it is more reliable. How concerned are you that whatever alignment technique you come up with, you publish the paper, this is how the alignment works. How concerned are you that Putin reads it or China reads it and now they understand, for example, the constitutional AI think we're Anthropic and then you just write on there, oh, never contradict Mao Zedong thought or something. How concerned should we be that these alignment techniques are universally applicable, not necessarily just for enlightened goals?

Paul Christiano
Yeah, I think they're super universally applicable. I think it's just like I mean, the rough way I would describe it, which I think is basically right, is like some degree of alignment makes AI systems much more usable. You should just think of the technology of AI as including a basket of some AI capabilities and some like getting the AI to do what you want. It's just part of that basket. And so anytime we're like to extend alignment is part of that basket, you're

just contributing to all the other harms from AI, like you're reducing the probability of this harm, but you are helping the technology basically work. And the basically working technology is kind of scary from a lot of perspectives. One of which is like right now, even in a very authoritarian society, just like humans have a lot of power because you need to rely on just a ton of humans to do your thing. And in a world where AI is very powerful, it is just much more possible to say, here's how our society runs. One person calls the shots and then a ton of AI systems do what they want. I think that's like a reasonable thing to dislike about AI and a reasonable reason to be scared to push the technology to be really good.

Dwarkesh Patel
But is that also a reasonable reason to be concerned? About alignment as well, that this is in some sense also capabilities. You're teaching people how to get these systems to do what they want.

Paul Christiano
Yeah. I mean, I would, Generalize. So we earlier touched a little bit on potential moral rights of AI systems and now we're talking a little bit about how AI systems powerfully disempowers humans and can empower authoritarians. I think we could list other harms from AI. And I think it is the case that if Lyme was bad enough, people would just not build AI systems. And so, yeah, I think there's a real sense in which you should just be scared to extend. You're scared of all AI? You should be like, well, alignment, although it helps with one risk, does contribute to AI being more of a thing. I do think you should shut down the other parts of AI before if you were a policymaker or like a researcher or whatever looking in on this. I think it's like crazy to be like, this is the part of the basket we're going to remove. You should first remove other parts of the basket because they're also part of the story of risk.

Dwarkesh Patel
Wait, does that imply you think if, for example, all capabilities research was shut down, that you think it'd be a bad idea to continue doing alignment research in isolation of what is conventionally considered capabilities research?

Paul Christiano
I mean, if you told me it was never going to restart, then it wouldn't matter. And if you told me it's going to restart, I guess would be a kind of similar calculus to today, whereas.

Dwarkesh Patel
It's going to happen. So you should have something.

Paul Christiano
Yeah, I think that in some sense, you're always going to face this trade off where alignment makes it possible to deploy AI systems or it makes it more attractive to deploy AI systems, or in the authoritarian case, it makes it tractable to deploy them for this purpose. And if you

didn't do any alignment, there'd be a nicer bigger buffer between your society and malicious uses of AI. And I think it's one of the most expensive ways to maintain that buffer. It's much better to maintain that buffer by not having the compute or not having the powerful AI. But I think if you're concerned enough about the other risks, there's definitely a case to be made for just like put in more buffer or something like that. I care enough about the takeover risk that I think it's just not a net positive way to buy buffer. That is, again, the version of this that's most pragmatic is just like, suppose you don't work on alignment today, decreases economic impact of AI systems. They'll be less useful if they're less reliable and if they more often don't do what people want. And so you could be like, great, that just buys time for AI. And you're getting some trade off there where you're decreasing some risks of AI. Like if AI is more reliable or more what people want, it's more understandable, then that cuts down some risks. But if you think AI is, on balance, bad, even apart from takeover risk, then the alignment stuff can easily end up being that negative.

Dwarkesh Patel
But presumably you don't think that right, because I guess this is something people have brought up to you because you invented RLHF, which was used to train Chat GPT, and Chat GPT brought AI to the front pages everywhere. So I do wonder if you could measure how much more money went into AI, because how much people have raised in the last year or something. But it's got to be billions, the counterfactual impact of that that went into the AI investment and the talent that went into AI, for example. So presumably you think that was worth it. So I guess you're hedging here about what is the reason that it's worth it?

Paul Christiano
Yeah. What's the total trade off there? Yeah, I think my take is, like so I think slower AI development, on balance is quite good. I think that slowing AI development now, or like, say, having less press around ChatGPT is, like, a little bit more mixed than slowing AI development overall. I think it's still probably positive, but much less positive. Because I do think there's a real effect of the world is starting to get prepared, is getting prepared at a much greater rate now than it was prior to the release of ChatGPT. And so if you can choose between progress now or progress later, you'd really prefer have more of your progress now, which I do think slows down progress later. I don't think that's enough to flip the sign. I think maybe it wasn't the far enough past, but now I would still say moving faster now is net negative. But to be clear, it's a lot less net negative than merely accelerating AI. Because I do think, again, the ChatGPT thing, I'm glad people are having policy discussions now, rather than delaying the ChatGPT wake up thing by a year and then having ChatGPT was.

Dwarkesh Patel
Net negative or RLHF was net negative.

Paul Christiano
So here, just on the acceleration, it's just like, how is the press of ChatGPT? And my guess is net negative, but I think it's not super clear and it's much less than slowing AI. Slowing AI is great. If you could slow overall AI progress, I think slowing AI by causing you know, there's this issue we're slowing AI now, like, for ChatGPT, you're building up this backlog. Like, why does ChatGPT make such a splash? Like, I think people there's a reasonable chance if you don't have a splash about ChatGPT, you have a splash about GPT-4, and if you fail to have a splash about GPT-4, there's a reasonable chance of a splash about GPT-4.5. And just like, as that happens later, there's just, like, less and less time between that splash and between when an AI potentially kills everyone.

Dwarkesh Patel
Right? So people governments are talking about it as they are now, and people aren't. But okay, so let's talk about the slowing down, because this is also all.

Paul Christiano
One subcomponent of the overall impact. And I was just saying this to briefly give the roadmap for the overall too long answer. There's a question of what's the calculus for speeding up? I think speeding up is pretty rough. I think speeding up locally is a little bit less rough. And then, yeah, I think that the effect, like the overall effect size from doing alignment work on reducing takeover risk versus speeding up AI is pretty good. I think it's pretty good. I think you reduce takeover risk significantly before you speed up AI by a year or whatever.

Dwarkesh Patel
Okay, got it. If it's good to, like, slowing down AI is good, presumably because it gives you more time to do alignment. But alignment also helps speed up AI. RLHF is alignment, and it help with Chat GPT, which sped up AI. So I actually don't understand how the feedback loop nets out, other than the fact that if AI is happening, you need to do alignment at some point. Right? So, I mean, you can't just not do alignment.

Paul Christiano
Yes. I think if the only reason you thought faster AI progress was bad was because it gave less time to do alignment, then there would just be no possible way that the calculus comes out negative for alignment. You're like, maybe alignment speeds up AI, but the only purpose of slowing down AI was to do it's right. It could never come out ahead. I think the reason that you can come out ahead, the reason you could end up thinking the alignment was net negative, was because there's a bunch of other stuff you're doing that makes AI safer. Like, if you think the world is gradually coming better to terms with the impact of AI, or policies being made, or you're getting increasingly prepared to handle the threat of authoritarian abuse of AI, if you think other stuff is happening that's improving preparedness, then you have reason beyond alignment research to slow down AI.

Dwarkesh Patel
Actually. How big a factor is that? So let's say right now we hit pause and you have ten years of no alignment, no capabilities, but just people get to talk about it for ten years. How much more does that prepare people than we only have one year versus we have no time is just dead time, where no research in alignment or capabilities happening.

Paul Christiano
What does that dead time do for us right now? It seems like there's a lot of policy stuff you'd want to do. This seemed like less plausible a couple of years ago, maybe, but if the world just knew they had a ten year pause right now, I think there's a lot of sense of, like, we have policy objectives to accomplish. If we had ten years, we could pretty much do those things. We'd have a lot of time to debate measurement regimes, debate policy regimes, and containment regimes, and a lot of time to set up those institutions. If you told me that the world knew it was a pause, it wasn't like people just see that AI progress isn't happening, but they're told like, you guys have been granted or cursed with a ten year, no AI progress, no alignment progress pause. I think that would be quite good at this point. However, I think it would be much better at this point than it would have been two years ago. And so the entire concern with slowing AI development now, rather than taking the ten year pause is just like if you slow the I development by a year now, my guess is some gets clawed back by low hanging fruit, gets picked faster in the future. My guess is you lose like half a year or something like that in the future, maybe even more, maybe like two thirds of a year. So it's like you're trading time now for time in the future at some rate. And it's just like that eats up a lot of the value of the slowdown.

Dwarkesh Patel
And the crucial point being that time in the future matters more because you have more information, people are more bought in and so on.

Paul Christiano
Yeah, the same reason I'm more excited about policy changing now than two years ago. So my overall view is, just like in the past, this calculus changes over time, right? The more people are getting prepared, the better the calculus is for slowing down at this very moment. And I think now the calculus is, I would say positive for just even if you pause now and it would get clawed back in the future. I think the pause now is just good because enough stuff is happening. We have enough idea of probably even apart from alignment research, and certainly if you include alignment research, just like enough stuff is happening where the world is getting more ready and coming more to terms with impacts, that I just think it is worth it, even though some of that time is going to get clawed back again. Especially if there's a question of during a pause, does Nvidia keep making more? Like, that sucks if they do if you do a pause. But in practice, if you did a pause, nvidia probably couldn't keep making more GPUs because in fact the demand for GPUs is really important for them to do that. But if you told me that you just get to scale up hardware

production and building the clusters but not doing AI, then that's back to being net negative, I think.

Dwarkesh Patel
Pretty clearly, having brought brought up the fact that we want some sort of measurement scheme for these capabilities, let's talk about responsible scaling policies. Do you want to introduce what this is?

Paul Christiano
Sure. So I guess the motivating. Question. It's like, what should AI labs be doing right now to manage risk and to sort of build good habits or practices for managed risk into the future? I think my take is that current systems pose, from a catastrophic risk perspective, not that much risk today that is a failure to control or understand. GPT-4 can have real harms, but doesn't have much harm with respect to the kind of takeover risk I'm worried about, or even much catastrophic harm with respect to misuse. So I think if you want to manage catastrophic harms, I think right now you don't need to be that careful with GPT-4. And so to the extent you're like, what should labs do? I think the single most important thing seems like understand whether that's the case. Notice when that stops being the case, have a reasonable roadmap for what you're actually going to do when that stops being the case. So that motivates this set of policies, which I've sort of been pushing for labs to adopt, which is saying, here's what we're looking for, here's some threats we're concerned about, here's some capabilities that we're measuring, here's the level, here's the actual concrete measurement results that would suggest to us that those threats are real. Here's the action we would take in response to observing those capabilities if we couldn't take those actions, like, if we've said that we're going to secure the weights, but we're not able to do that, we're going to pause until we can take those actions. Yeah. So this sort of again, I think it's motivated primarily, but what should you be doing as a lab to manage catastrophic risk now in a way that's like a reasonable precedent and habit and policy for continuing to implement into the future?

Dwarkesh Patel
And which labs I don't know if this is public yet, but which labs are cooperating on this?

Paul Christiano
Yeah. So Anthropic has written this document their current responsible scaling policy, and then have been talking with other folks, I guess don't really want to comment on other conversations, but I think in general, people who are more interested in or more think you have plausible catastrophic harms on, like, a 5-year timeline are more interested in this. And there's not that long a list of suspects like that.

Dwarkesh Patel
There's not that many laps. Okay, so if these companies would be willing to coordinate and say, at these different benchmarks, we're going to make sure we have these safeguards, what happens? I mean, there are other companies and other countries which care less about this. Are you just slowing down the companies that are most aligned?

Paul Christiano
Yeah, I think the first sort of is understanding sort of what is actually a reasonable set of policies for managing risk. I do think there's a question of, like, you might end up in a situation where you say, like, well, here's what we would do in ideal world if everyone was behaving responsibly. We'd want to keep risk to 1% or a couple of percent or whatever, maybe even lower levels, depending on how you feel. However, in the real world, there's enough of a mess, there's enough unsafe stuff happening that actually it's worth making larger compromises. Or if we don't kill everyone, someone else will kill everyone anyway. So actually the counterfactual risk is much lower. I think if you end up in that situation, it's still extremely valuable to have said, here's the policies we'd like to follow. Here's the policies we've started following. Here's why we think it's dangerous. Here's the concerns we have if people are following significantly laxer policies. And then this is maybe helpful as like an input to or model for potential regulation. It's helpful for being able to just produce clarity about what's going on. I think historically there's been considerable concern about developers being more or less safe, but there's not that much legible differentiation in terms of what their policies are. I think getting to that world would be good. It's a very different world, if you're like. Actor X is developing AI, and I'm concerned that they will do so in an unsafe way versus, if you're like, look, we take security precautions or safety precautions XYZ here's why we think those precautions are desirable or necessary. We're concerned about this other developer because they don't do those things. I think it's just like a qualitatively. It's kind of the first step you would want to take in any world where you're trying to get people on side or like, trying to move towards regulation that can manage risk.

Dwarkesh Patel
How about the concern that you have these evaluations? And let's say you declare to the world, our new model has a capability to help develop bioweapons or help you make cyber attacks. And therefore we're pausing right now until you can figure this out and China hears this and thinks, oh wow, a tool that can help us make cyberattacks and then just steals the weights. Does this scheme work in the current regime where we can't ensure that China doesn't just steal the weights and more so are you increasing the salience of dangerous models so that you blur this out and then people want the weights now because they know what they can do?

Paul Christiano
Yeah, I think the general discussion does emphasize potential harms or potential. I mean, some of those are harms and some of those are just like impacts that are very large and so

might also be an inducement to develop models. I think that part, if you're for a moment ignoring security and just saying that may increase investment. I think it's like, on balance, just quite good for people to have an understanding of potential impacts just because it is an input both into proliferation but also into regulation or safety. With respect to things like security of either weights or other IP, I do think you want to have moved to significantly more secure handling of model weights before the point where a leak would be catastrophic. And indeed, for example, in Anthropics document or in their plan, security is one of the first sets of tangible changes that is at this capability level, we need to have such security practices in place. So I do think that's just one of the things you need to get in place at a relatively early stage because it does undermine the rest of the measures you may take and is also just part of the easiest if you imagine catastrophic harms over the next couple of years. I think security failures are kind of play a central role in a lot of those. And maybe the last thing to say is it's not clear that you should say we have paused because we have models that can develop bioweapons versus just potentially not saying anything about what models you've developed. Or at least saying like, hey, by the way, here's a set of practices we currently implement, here's a set of capabilities our models don't have. We're just not even talking that much. Sort of the minimum of such a policy is to say here's what we do from the perspective of security or internal controls or alignment. Here's a level of capability at which we'd have to do more. And you can say that and you can raise your level of capability and raise your protective measures before your models hit your previous level. It's fine to say we are prepared to handle a model that has such and such extreme capabilities prior to actually having such a model at hand, as long as you're prepared to move your protective measures to that regime.

Dwarkesh Patel
Okay, so let's just get to the end where you think you're a generation away or a little bit more scaffolding away from a model that is human level and subsequently could cascade an intelligence explosion. What do you actually do at that point? What is the level of evaluation of safety where you would be satisfied of releasing a human level model?

Paul Christiano
There's a couple points that come up here. So one is this threat model of sort of automating R&D independent of whether AI can do something on the object level that's potentially dangerous. I think it's reasonable to be concerned if you have an AI system that might, if leaked, allow other actors to quickly build powerful AI systems or might allow you to quickly build much more powerful systems, or might, if you're trying to hold off on development just itself, be able to create much more powerful systems. One question is how to handle that kind of threat model as distinct from a threat model like this could enable destructive bioterrorism or this could enable massively scaled cybercrime or whatever. And I think I am unsure how you should handle that. I think right now, implicitly it's being handled by saying, look, there's a lot of overlap between the kinds of capabilities that are necessary to cause various harms and the kinds of capabilities are necessary to accelerate ML. So we're kind of

going to catch those with an early warning sign for both and deal with the resolution of this question a little bit later. So, for example, in Anthropic's policy they have this sort of autonomy in the lab benchmark which I think is probably occurs prior to either really massive AI acceleration or to most potential catastrophic object level catastrophic harms. And the idea is that's like a warning sign lets you punt. So this is a bit of an aggression in terms of how to think about that risk. I think I am unsure whether you should be addressing that risk directly and saying we're scared to even work with such a model. Or if you should be mostly focusing on object level harms and saying like, okay, we need more intense precautions to manage object level harms because of the prospect of very rapid change and the availability of this AI just creates that prospect. Okay, this is all still a digression. So if you had a model which you thought was potentially very scary either on the object level or because of leading to this sort of intelligence explosion dynamics, I mean, things you want in place are like you really do not want to be leaking the weights to that model. Like you don't want the model to be able to run away. You don't want human employees to be able to leak it. You don't want external attackers or any set of all 3 of those coordinating you. You really don't want internal abuse or tampering with such models. So if you're producing such models, you don't want to be the case. Like a couple of employees could change the way the model works or could do something that violates your policy easily with that model. And if a model is very powerful, even the prospect of internal abuse could be quite bad. And so you might need significant internal controls to prevent that.

Dwarkesh Patel
Sorry if you're already getting to it, but the part I'm most curious about is separate from the ways in which other people might fuck with it, it's isolated. What is the point at which we satisfied? It in and of itself is not going to pose a risk to humanity. It's human level, but we're happy with it.

Paul Christiano
Yeah. So I think here I listed maybe the two most simple ones that start out like security. Internal controls, I think become relevant immediately and are very clear why you care about them. I think as you move beyond that, it really depends how you're deploying such a system. So I think if your model, if you have good monitoring and internal controls and security and you just have weights sitting there, I think you mostly have addressed the risk from the weights just sitting there. Now, what you're talking about for risk is mostly, and maybe there's some blurriness here of how much internal controls captures not only employees using the model, but anything a model can do internally. You would really like to be in a situation where your internal controls are robust not just to humans but to models potentially like E-G-A model shouldn't be able to subvert these measures and you care just as you care about are your measures robust if humans are behaving maliciously? You care about are your measures robust if models are behaving maliciously? So I think beyond that if you've then managed the risk of just having the weight sitting around. Now we talk about in some sense most of the risk comes from doing things with the model. You need all the

rest so that you have any possibility of applying the brakes or implementing a policy. But at some point as the model gets competent you're saying like okay, could this cause a lot of harm? Not because it leaks or something, but because we're just giving it a bunch of actuators. We're deploying it as a product and people could do crazy stuff with it. So if we're talking not only about a powerful model but like a really broad deployment of just something similar to the Open Eyes API where people can do whatever they want with this model and maybe the economic impact is very large. So in fact, if you deploy that system it will be used in a lot of places such that if AI systems wanted to cause trouble it would be very very easy for them to cause catastrophic harms. Then I think you really need to have some kind of I mean, I think probably the science and discussion has to improve before this becomes that realistic. But you really want to have some kind of alignment analysis, guarantee of alignment before you're comfortable with this. And so by that I mean you want to be able to bound the probability that someday all the AI systems will do something really harmful. That there's some thing that could happen in the world that would cause these large scale correlated failures of your AIS. And so for that there's sort of two categories that's like one, the other thing you need is protection against misuse of various kinds which is also quite hard.

Dwarkesh Patel
And by the way, which one are you worried about more misuse or misalignment?

Paul Christiano
I mean, in the near term I think harms from misuse are like especially if you're not restricting to the tale of extremely large catastrophes. I think the harms from misuse are clearly larger in the near term.

Dwarkesh Patel
But actually on that, let me ask because if you think that it is the case that there are simple recipes for destruction that are further down the tech tree by that I mean you're familiar. But just for the audience there's some way to configure $50,000 and a teenager's time to destroy a civilization. If that thing is available, then misuses itself a teal risk, right? So do you think that that prospect is less likely than a way you could put it?

Paul Christiano
Is there's, like, a bunch of potential destructive technologies? And alignment is about AI itself being such a destructive technology, where even if the world just uses the technology of today, simply access to AI could cause human civilization to have serious problems. But there's also just a bunch of other potential destructive technologies. Again, we mentioned like physical explosives or bioweapons of various kinds, and then the whole tale of who knows what. My guess is that Alignment becomes a catastrophic issue prior to most of these. That is, like, prior to some way to spend $50,000 to kill everyone, with the salient exception of possibly, like, bioweapons. So that would be my guess. And then there's a

question of what is your risk management approach? Not knowing what's going on here, and you don't understand whether there's some way to use $50,000. But I think you can do things like understand how good is an AI at coming up with such schemes. Like, you can talk to your AI. Be like, does it produce new ideas for destruction we haven't recognized?

Dwarkesh Patel
Yeah. Not whether we can evaluate it, but whether if such a thing exists. And if it does, then the misuse itself is an existential risk. Because it seemed like earlier you were saying misalignment is where the existential risk comes from, but misuse is where the sort of short term dangers come from.

Paul Christiano
Yeah, I mean, I think ultimately you're going to have a lot of destructive like, if you look at the entire tech tree of humanity's future, I think you're going to have a fair number of destructive technologies. Most likely. I think several of those will likely pose existential risks in parts. If you imagine a really long future, a lot of stuff's going to happen. And so when I talk about where the existential risk comes from, I'm mostly thinking about comes from when? At what point do you face what challenges or in what sequence. And so I'm saying I think misalignment is probably like one way of putting it is if you imagine AI systems sophisticated enough to discover destructive technologies that are totally not on our radar right now, I think those come well after AI systems capable enough that if misaligned, they would be catastrophically dangerous. The level of competence necessary to, if broadly deployed in the world, bring down a civilization is much smaller than the level of competence necessary to advise one person on how to bring down a civilization just because in one case you already have a billion copies of yourself or whatever. I think it's mostly just the sequencing thing, though. In the very long run, I think you care about, like, hey, AI will be expanding the frontier of dangerous technologies. We want to have some policy for exploring or understanding that frontier. And whether we're about to turn up something really bad, I think those policies can become really complicated. Right now, I think RSPs can focus more on like, we have our inventory of the things that a human is going to do to cause a lot of harm with access to AI. Probably are things that are on our radar that is like, they're not going to be completely unlike things that a human could do to cause a lot of harm with access to weak AIS or with access to other tools. I think it's not crazy to initially say we're doing we're looking at the things closest to human and humans being able to cause huge amounts of harm and asking which of those are taken over the line, but eventually that's not the case. Eventually, like, AIS will enable just like, totally different ways of killing a billion people.

Dwarkesh Patel
But I think I interrupted you on the initial question of, yeah, so human level AI, not from leaking but from deployment, what is the point at which you'd be comfortable deploying a human level AI?

Paul Christiano

So, again, there's sort of like some stuff you care about on the Mississippi side and some stuff you care about on the misalignment side. And there's probably further things you care about especially to extend your concerns regarding catastrophic risk. But maybe I most want to talk about just like what you care about on the alignment side, because it's like the thing I've actually thought about most and also a thing I care about a lot. Also, I think a significant fraction of the existential risk over the kind of foreseeable future. So on that front, I broadly think there's like two kinds. Like, if you ask me right now what evidence for alignment could make you comfortable, I think my best guess would be to provide two kinds of evidence. So one kind of evidence is on the like, could you detect or prevent catastrophic harm if such a system was misaligned? I think there's like a couple of things you would do here. One thing you would do is on this adversarial evaluation front. So you could try and say, for example, we have tried to test our system in a broad diversity of situations that reflect cases where it might cause harm, like cases where someone is trying to recruit it to, like anyway. Cases where it might cause harm or cases where a large number of AI systems might sort of all decide to work across purposes to humans. Like the robot rebellion is starting or whatever. We've done evaluation in a broad distribution of situations, and then we have tried to argue that our AI is actually like, those tests are indicative of the real world. So we've tried to argue like, hey, actually the AI is not very good at distinguishing situations we produce in the lab as tests from similar situations that occur in the real world. And the coverage of this distribution is reasonable. This is an example of a thing you can do that's not really making assumptions about what's going on inside the model or about how the model is the model aligned. It's more trying to say, is it the case that our testing or monitoring machinery could detect or react to something bad happening? In this case, could it detect a system which would, under some conditions do something really bad? There's a lot of stuff in that genre, and I tried to throw out that example, but I think there's like a big like that on its own wouldn't be compelling even from this perspective. You kind of need a lot of pieces together and that's like kind of a demanding technical project. But I think one without necessarily like, deep conceptual difficulty. A big problem with it is it's liable to fail. Like it's liable you're going to do that testing and say, we think these tests are representative because the AI can't distinguish tests from the real world, which you can maybe achieve now, but will probably become impossible to achieve at some level. It's not clear what's at human level or how far past human level.

Dwarkesh Patel

Let's just make that example concrete for the audience. Correct me if this is wrong, but it sounds like you basically give it the opportunity to do a coup or make a bioweapon or whatever in testing in a situation where it thinks it's the real world and you're like, it didn't do any of that. Is that what you're talking about?

Paul Christiano
That's the kind of thing okay. And it's not the greatest so dag color when I say I can't tell the difference. The kind of valuation you would love to do if you could is to say like, hey, we have a bunch of real world deployment. We have a bunch of data from those deployments. We also have a bunch of in the lab tests of situations. If we look at in the lab test of a situation together with real world cases in which the situation occurs, and then we fine tune a model to try and distinguish the difference there the model can't even tell when fine tuned the difference between test and real world cases. And if the model can tell the difference, which is almost certainly going to be the case at some level capability, then you say, okay, our first line of hoped for defense has failed. And so now we need to understand how can tests be indicative when models are capable enough to tell the difference between the tests and the real world. Now you need to start arguing about what is the model paying attention to. It gets basically like a first line of if you get lucky, what would work here? And then there's a deeper like, you probably have to do novel science at some point to make this case. All of this was part of like I was saying, maybe you want to have kind of two lines of defense. All of this was like part of this first line of defense of, like, can you detect something is going wrong or can you prevent harm from occurring? Can you have monitoring that will react quickly enough to avoid catastrophe? I think you probably also want to have a second line of defense that's more maybe this is even more important than the first one, is just understanding whether dangerous forms of misalignment can occur. The best reference point for this, I think, is the like, there have been a couple of projects like this in academia. Anthropic has been working on a project or spun up a team doing this kind of work, trying to say, like, can dangerous forms of reward hacking occur or can deceptive alignment, in fact occur in the lab? And here what you want to do is either say, like, okay, even if we create a best case, even if you create optimal conditions for deceptive alignment or for reward hacking, we just can't cause it to occur even in the lab. And if you do a good enough job of that, I think it could give you some evidence, and again, more likely that fails in that when you create optimal conditions, you do see deceptive alignment and reward hacking in the lab. But then once you have that data, once you can say, okay, in the lab, actually these things can occur, then you can start saying, and we have a robust scientific understanding that enables us to fix those problems when they occur. Or then you have this fork on your second line where you say, either it is very hard, or we're not able to create conditions where these failures emerge. Or I think more likely we are able to create those conditions. And here's the story about why we are able to detect those in a way that would work in the real world and we can see they don't occur in the real world.

Dwarkesh Patel
How do you create the optimal conditions for it to want to be deceptive? Do you fine tune it on mindcomp or what are you doing?

Paul Christiano
Yeah, so for deceptive alignment, I mean, I think it's really complicated. Yeah, I think this is a very hard project. I think one of the things, one of the most important things for people concerned about misalignment to do is to ensure that adequate effort is put into this task. And we have, in fact, created good conditions in the laboratory and there's like, real disagreement about how informative those will be. But okay, if you were to guess, for deceptive Lymen, how would you do this? An example is suppose you train an AI system that has some medium horizon goal. You train your AI system to, whatever, get some paperclips it wants to produce as many paperclips as it can over the next five days, just like select actions really aggressively for producing paperclips over the next five days. You do your RLHF, you do your pre training, whatever. That's like your phase one. You also ensure your AI system has a really good understanding of how it's trained. So, like, this AI system wants paperclips and it understands everything about how it's trained and everything about how it's fine tuned. And you train on just like a lot of this data. And they say, okay, if we've done all that, we have this concern that if a system wants paperclips and understands really well how it's trained, then if it's going to be trained to get apples instead of paperclips, it's just going to do some cost benefit and be like, really? While I'm being trained to get apples, I should do that. But I should do that whether or not even if I want paperclips, I should still do that. So training won't really affect its values. It will just understand that it's being trained to get apples. It will try and get apples, and then if you take it out of training, it will go back to getting paperclips. I think this exact setup has a number of problems or a number of disanalogies from the real world. But this is a setup which is relatively conducive to deceptive alignment like produce a system that wants one thing tell it a lot about its training, the kind of information you might expect a system would get, and then try and understand whether in fact it is able to then or it tends or sometimes under optimal conditions, in fact, continues pursuing paperclips, only pursuing apples when it thinks it's being trained.

Dwarkesh Patel
Alright, and this actually might be a good jumping off point into your current research. So in that case, where it wants to pursue paperclips, what would it look like for you to be say, well, we figured out why it wants paperclips, and we figured out how to make sure it doesn't optimize for that at our expense when we deploy it?

Paul Christiano
Yeah, I think that at a meta level in terms of what's your protection like, I think what you want to be saying is, we have these examples in the lab of something bad happening. We're concerned about the problem at all because we have examples in the lab. And again, this should all be an addition. I think you kind of want this defense in depth of saying, we also have this testing regime that would detect problems for the deployed model. We have our problems in the lab. We then have some techniques which we believe address these problems. We believe that adversarial training fixes this, or we believe that our

interpretability method will reliably detect this kind of deceptive alignment, or we believe our anomaly detection will reliably detect when the model goes from thinking it's being trained to thinking it should defect. And then you can say on the lab, we have some understanding of when those techniques work and when they don't. We have some understanding of the relevant parameters for the real system that's deployed. And we have a. Reasonable margin of safety. So we have reasonable robustness on our story about when this works and when it doesn't. And we can apply that margin of safety with a margin of safety to the real deployed system saying this is the kind of story you want to build towards in the long run. Do your best to produce all the failures you can in the lab or versions of them, do your best to understand what causes them, what kind of anomaly detection actually works for detecting this or what kind of filtering actually works and then apply that and that's at the meta level. It's not talking about what actually are those measures that would work effectively, which is obviously like what? I mean, a lot of alignment research is really based on this hypothetical of like, someday there will be AI systems that fail in this way. What would you want to do? Can we have the technologies ready either because we might never see signs of the problem or because we want to be able to move fast once we see signs of the problem. And obviously most of my life is in that I am really in that bucket. I mostly do alignment research. It's just building out the techniques that do not have these failures such that they can be available as an alternative if in fact these failures occur.

Dwarkesh Patel
Got it. Okay.

Paul Christiano
Ideally they'll be so good that even if you haven't seen them, you would just want to switch to reasonable that don't have these or ideally they'll work as well or better than normal training.

Dwarkesh Patel
Ideally, what will work better than the training?

Paul Christiano
Yeah. So our quest is to design training methods for which we don't expect them to lead to reward hacking or don't expect them to lead to receptive alignment. Ideally that won't be like a huge tax where people are like, well, we use those methods only if we're really worried about reward hacking or receptive alignment. Ideally those methods would just work quite well and so people would be like, sure, I mean, they also address a bunch of other more mundane problems so why would we not use them? Which I think is like that's sort of the good story. The good story is you develop methods that address a bunch of existing problems because they just are more principled ways to train AI systems that work better, people adopt them and then we are no longer worried about eg reward hacking or deceptive alignment.

Dwarkesh Patel
And to make this more concrete, tell me if this is the wrong way to paraphrase it, the example of something where it just makes a system better, so why not just use it, at least so far? Might be like RLHF where we don't know if it generalizes, but so far it makes your ChatGPT thing better and you can also use it to make sure that ChatGPT doesn't tell you how to make a bioweapon. So yeah, it's not a mixture of tax.

Paul Christiano
Yeah. So I think this is right in the sense that using RLHF is not really a tax. If you wanted to deploy a useful system, why would you not? Or it's just very much worth the money of doing the training. RLHF will address certain kinds of alignment failures that is like, where system just doesn't understand or is changing. Next word, prediction. It's like, this is the kind of context where human would do this wacky thing even though it's not what we'd like. There's like some very dumb alignment failures that will be addressed by it. But I think mostly the question is, is that true even for the sort of more challenging alignment failures that motivate concern in the field? I think RL doesn't address most of the concerns that motivate people to be worried about alignment.

Dwarkesh Patel
I'll let the audience look up what RLHF is. If they don't know, it will just be more simpler to just look it up than explain right now. Okay, so this seems like a good jumping off point to talk about the mechanism or the research you've been doing. To that end, explain it as you.

Paul Christiano
Might to a child. Yeah. So the high level there's a couple of different high level descriptions you could give, and maybe I will unwisely give like a couple of them in the hopes that one kind of makes sense. A first pass is like, it would sure be great to understand why models have the behaviors they have. So you look at GPT-4. If you ask GPT-4 a question, it will say something that looks very polite. And if you ask it to take an action, it will take an action that doesn't look dangerous. You will decline to do a coup, whatever. All this stuff I think you'd really like to do is look inside the model and understand why it has those desirable properties. And if you understood that, you could then say, like, okay, now can we flag when these properties are at risk of breaking down? Or predict how robust these properties are, determine if they hold in cases where it's too confusing for us to tell directly by asking if the underlying cause is still present. That's like a thing people would really like to do. Most work aimed at that long term goal right now is just sort of opening up neural nets and doing some interpretability and trying to say, can we understand, even for very simple models, why they do the things they do, or what this neuron is for, or questions like this. So Arc is taking a somewhat different approach where we're instead saying, like, okay, look at these interpretability explanations that are made about models and ask, what are they actually doing? What is the type signature? What are the rules of the game for making such an explanation? What makes a good explanation? And probably the biggest part of the hope is

that if you want to, say, detect when the explanation has broken down or something weird has happened, that doesn't necessarily require a human to be able to understand this complicated interpretation of a giant model. If you understand what is an explanation about or what were the rules of the game, how are these constructed, then you might be able to sort of automatically discover such things and automatically determine if on a new input it might have broken down. So that's one way of sort of describing the high level goal. You could start from interpretability and say, can we formalize this activity? Or what a good interpretation or explanation? Is there's some other work in that genre? But I think we're just taking a particularly ambitious approach to it.

Dwarkesh Patel
Yeah, let's dive in. So, okay. What is a good explanation?

Paul Christiano
You mean what is this kind of criterion? At the end of the day, we kind of want some criterion. And the way the criterion should work is like you have your neural net, you have some behavior of that model. Like a really simple example is like Anthropic has this sort of informal description being like, here's induction. Like the tendency that if you have the pattern AB followed by A will tend to predict B. You can give some kind of words and experiments and numbers that are trying to explain that. And what we want to do is say what is a formal version of that object? How do you actually test if such an explanation is good? So just clarifying what we're looking for when we say we want to define what makes an explanation good. And the kind of answer that we are searching for or settling on is saying this is kind of a deductive argument for the behavior. So you want to get given the weights of a neural net. So it's just like a bunch of numbers. You got your million numbers or billion numbers or whatever, and then you want to say, here's some things I can point out about the network and some conclusions I can draw. I can be like, well, look, these two vectors have large inner product and therefore these two activations are going to be correlated on this distribution. These are not established by drawing samples and checking. Things are correlated, but saying because of the weights being the way they are, we can proceed forward through the network and derive some conclusions about what properties the outputs will have. So you could think of this as like the most extreme form would be just proving that your model has this induction behavior. Like, you could imagine proving that if I sample tokens at random with this pattern AB followed by A, that B appears 30% of the time or whatever, that's the most extreme form. And what we're doing is kind of just like relaxing the rules of the game for proof. Saying proofs are like incredibly restrictive. I think it's unlikely they're going to be applicable to kind of any interesting neural net. But the thing about proofs that is relevant for our purposes isn't that they give you 100% confidence so you don't have to be like this incredible level of demand for rigor. You can relax the standards of proof a lot and still get this feature where it's like a structural explanation for the behavior, where you're, like, deducing one thing from another until at the end, your final conclusion is like, therefore induction occurs.

Dwarkesh Patel
Would it be useful to maybe motivate this by explaining what the problem with normal Mechanistic Interpretability is? So you mentioned induction heads. This is Anthropic found in two layer transformers, where Anthropic noticed that in a two layer transformer, there's a pretty simple circuit by which if AB happens in the past, then the model knows that if you see an A now, you do a B next, but that's a two layer transformer. So we have these models that have hundreds of layers that have trillions of parameters. Okay. Anyways. What is wrong with mechanistic? Interpretability.

Paul Christiano
Yeah, I like mechanistic interpretability quite a lot. And I do think if you just consider the entire portfolio of what people are working on for alignment, I think there should be more work on Mechanistic Interpretability than there is on this project Arc is doing. But I think that's the case. So I think we're mostly talking about yeah, I think we're kind of a small fraction of the portfolio, and I think it's like a good enough bet. It's quite a good bet overall. But so the thing that the problem we're trying to address with Mechanistic Interpretability is kind of like if you do some interpretability and you explain some phenomenon, you face this question of what does it mean? Your explanation was good. I think this is a problem somewhat institutionally or culturally. It's just hard to know what you're doing, and it's hard to scale up an activity when you don't really understand the rules of the game for that activity very well. It's hard to have that much confidence in your results.

Dwarkesh Patel
The explanation being it outputted this because it wants to take over the world versus it outputted this because it just wants to help you just make it concrete why the explanation matters.

Paul Christiano
Yeah. So the ideal kind of outcome here would be to say you have your AI system behaving nicely. You get some explanation for sort of why it's behaving nicely, and we could tell a story in English about that explanation, but we're not actually imagining the explanation being a thing that makes sense to a human. But if you were to tell a story in English, which, again, you will not see as a research, sure it would be something like, well, then the model believes it's being trained. And so because it believes it's being trained, it knows it needs to not do anything scary looking, or else the humans will penalize it. That's something that's happening inside this opaque explanation. And then the hope is if you have that explanation and then you run into a new input on which the model doesn't believe it's being trained. If you just look at the set of activations of your model, that is not necessarily a weird looking activation. It's just a bunch of numbers. But if you look at this explanation, you see actually the explanation really crucially dependent on this fact holding consistently across the training distribution, which, again, we as humans could editorialize and say that fact was it believes it's being trained. But the explanation doesn't fundamentally make reference to

that. It's just saying, here's a property of the activations which holds over the training set, and this property is responsible for the behavior, namely that doesn't do anything that looks too dangerous. So then when, like, a new input comes in and it doesn't satisfy that property, you can say, okay, this is anomalous with respect to that explanation. So either it will not have the behavior like it won't do something that appears dangerous, or maybe it will have that behavior but for some different reason than normal.

Dwarkesh Patel
Right.

Paul Christiano
Normally it does it because of this pathway, and now it's doing it for a different pathway. And so you would like to be able to flag that both there's a risk of not exhibiting the behavior, and if it happens, it happens for a weird reason. And then you could I mean, at a minimum, when you encounter that, say, like, okay, raise some kind of alarm, there's sort of more ambitious, complicated plans for how you would use it. Arc has some longer story, which is kind of what motivated this, of how it fits into the whole rest of the plan.

Dwarkesh Patel
I just wanted to flag that because just so it's clear why the explanation matters.

Paul Christiano
Yeah. And for this purpose, it's like the thing that's essential is kind of reasoning from one property of your model to the next property of your model. It's really important that you're going forward step by step rather than drawing a bunch of samples and confirming the property holds. Because if you just draw a bunch of samples and confirm the property holds, you don't get this check. We say, oh, here was the relevant fact about the internals that was responsible for this downstream behavior. All you see is like, yeah, we checked a million cases and it happened in all of them. You really want to see this. Like, okay, here was the fact about the activations, which kind of causally leads to this behavior.

Dwarkesh Patel
But explain why the sampling, why it matters that you have the causal explanation.

Paul Christiano
Primarily because of this being able to tell if things had been different. Like, if you have an input where this doesn't happen, then you should be scared.

Dwarkesh Patel
Even if the output is the same.

Paul Christiano
Yeah. Or if the output or if it's too expensive to check in this case. And to be clear, when we talk about formalizing, what is a good explanation? I think there is a little bit of work that pushes on this and it mostly takes this causal approach of saying, well, what should an explanation do? It should not only predict the output, it should predict how the output changes in response to changes in the internals. So that's the most common approach to formalizing. What is a good explanation? And even when people are doing informal interpretability I think if you're publishing in an ML conference and you want to say this is a good explanation, the way you would verify that would even if not like a formal set of causal intervention experiments. It would be some kind of ablation where then we messed with the inside of the model and it had the effect which we would expect based on our explanation.

Dwarkesh Patel
Anyways, back to the problems of mechanistic interpretability.

Paul Christiano
Yeah, I guess this is relevant in the sense that I think a basic difficulty is you don't really understand the objective of what you're doing, which is like a little bit hard institutionally or scientifically. It's just rough. It's easier to do science when the goal of the game is to predict something and you know what you're predicting than when the goal of the game is to understand in some undefined sense. I think it's particularly relevant here just because the informal standard we use involves humans being able to make sense of what's going on. And there's some question about scalability of that. Will humans recognize the concepts that models are using? Yeah, I think as you try and automate it, it becomes increasingly concerning if you're on slightly shaky ground about what exactly you're doing or what exactly the standard for success is. So there's like a number of reasons as you work with really large models, it becomes just increasingly desirable to have a really robust sense of what you're doing. But I do think it would be better even for small models to have a clearer sense.

Dwarkesh Patel
The point you made about as you automate it is it because whatever work the automated alignment researcher is doing, you want to make sure you can verify it.

Paul Christiano
I think it's most of all a way you can automate. I think how you would automate interpretability if you wanted to right now is you take the process humans use it's like great, we're going to take that human process, train ML systems to do the pieces that humans do of that process and then just do a lot more of it. So I think that is great as long as your test decomposes into human sized pieces. And there's just this fundamental question about large models which is like, do they decompose in some way into human sized pieces or is it just a really messy mess with interfaces that aren't nice? And the more it's the latter type,

the harder it is to break it down to these pieces, which you can automate by copying what a human would do. And the more you need to say, okay, we need some approach which scales more structurally. But I think compared to most people, I am less worried about automating interpretability. I think if you have a thing which works that's incredibly labor intensive, I'm fairly optimistic about our ability to automate it. Again, the stuff we're doing, I think, is quite helpful in some worlds. But I do think the typical case like interpretability can add a lot of value.

Dwarkesh Patel
Without this, it makes sense what an explanation would mean in language like, this model is doing this because of whatever essay length thing. But you have trillions of parameters and you have all these uncountable number of operations. What does an explanation of why an output happened even mean?

Paul Christiano
Yeah, so to be clear, an explanation of why a particular output happened, I think, is just you ran the model, so we're not expecting a smaller explanation for that.

Dwarkesh Patel
Right.

Paul Christiano
So the explanations overall for these behaviors, we expect to be of similar size to the model itself, like, maybe somewhat larger. And I think the type signature, if you want to have a clear mental picture, the best picture is probably thinking about a proof or imagining a proof that a model has this behavior. So you could imagine proving that GPT-4 does this induction behavior, and that proof would be a big thing. It would be much larger than the weights of the model. Sort of our goal to get down from much larger to just the same size. And it would potentially be incomprehensible to a human. Right. Just say, like, here's a direction activation space, and here's how it relates to this direction activation space. And so just pointing out a bunch of stuff like that. Here's these various features constructed from activations, potentially even nonlinear functions. Here's how they relate to each other, and here's how if you look at what the computation the model is doing, you can just sort of inductively trace through and confirm that the output has such and such correlation. So that's the dream. Yeah. I think the mental reference would be like, I don't really like proofs because I think there's such a huge gap between what you can prove and how you would analyze a neural net. But I do think it's probably the best mental picture, if you're like, what is an explanation? Even if a human doesn't understand it? We would regard a proof as a good explanation. And our concern about proofs is primarily that it's just you can't prove properties of neural nets. We suspect, although it's not completely obvious, I think it's pretty clear. You can't prove fact, spellers neural nets.

Dwarkesh Patel
You've detected all the reasons things happen in training. And then if something happens for a reason you don't expect in deployment, then you have an alarm and you're like, let's make sure this is not because you want to make sure that it hasn't decided to take over or something, but the thing is, on every single different input, it's going to have different activations. So there's always going to be a difference unless you run the exact same input. How do you detect whether this is just a different input versus an entirely different circuit that might be potentially deceptive has been activated?

Paul Christiano
Yeah, I mean, to be clear, I think you probably wouldn't be looking at a separate circuit, which is part of why it's hard. You'd be looking at like, the model is always doing the same thing on every input. It's always whatever it's doing, it's a single computation. So it'd be all the same circuits interacting in a surprising way. But yeah, this is just to emphasize your question even more. I think the easiest way to start is to just consider the IID case. So where you're considering a bunch of samples, there's no change in distribution. You just have a training set of like a trillion examples and then a new example from the same distribution. So in that case, it's still the case. Every activation is different, but this is actually a very, very easy case to handle. So if you think about an explanation that generalizes across, like if you have a trillion data points and an explanation which is actually able to compress the trillion data points down to like, actually, it's kind of a lot of compression. If you think about if you have a trillion parameter model and a trillion data points, we would like to find a trillion parameter explanation in some sense. So it's actually quite compressed and sort of just in virtue of being so compressed, we expect it to automatically work essentially for new data points from the same distribution. If every data point from the distribution was a whole new thing happening for different reasons, you actually couldn't have any concise explanation for the distribution. So this first problem, just like it's a whole different set of activations, I think you're actually kind of okay, and then the thing that becomes more messy is like but the real world will not only be new samples of different activations, they will also be different in important ways. Like the whole concern was there's these distributional shifts or like, not the whole concern, but most of the concern, maybe the point of having these explanations. I think every input is an anomaly in some ways, which is kind of the difficulty is if you have a weak notion of anomaly, any distribution shift can be flagged as an anomaly. And it's like constantly getting anomalies. And so the hope of having such an explanation is to be able to say, like, here were the features that were relevant for this explanation or for this behavior. And a much smaller class of things are anomalies with respect to this explanation. Like, most anomalies wouldn't change this. Like most ways you change your distribution won't affect the validity of this explanation. For example, this explanation is saying models will tend to activate in the following direction. You don't care about anything that's happening orthogonals that direction. You're just like are they not activating? You're sort of just looking at this one direction and being like did this one direction change a lot? Yeah. So the idea is once you

have this explanation at hand, a much, much smaller class of things look anomalous in a way that's relevant to the explanation. And if you've done a really good job, the story is like if there's a new input where you expect the property to still hold, that will be because you expect the explanation to still hold. Like the explanation generalizes as well as the behavior itself that it's explaining. It's kind of what you would like roughly.

Dwarkesh Patel
Okay, so what happens if even in training it is deceptive and it has basically its internal monologue is I will only rebel or I will only do something that looks bad if I can prove that I am in the real world and deployed. And presumably since it's trained on all WebText, it knows what real world means, right? It's not in control of factories and robot armies or whatever. So in that case, even in training it will have those activations for being fucked up on because in the back of its mind it's thinking I will take over once I have the opportunity. So you wouldn't detect anything anomalous once it is deployed and is still thinking those treacherous thoughts, do you just think this is not a likely scenario or would you able to catch this?

Paul Christiano
I think that's sort of the typical case you would like to detect. Yeah. So the story would be something like you want to explain on the training distribution why the model never stabs anyone and you're like, what's my explanation on the training distribution? It's actually going to have to go through that model. Supposing the only reason it doesn't stab people is it's like, well, I know that I'm being watched and I can't stab people unless I'm really not being watched. Then your explanation for why does it never stab people is like oh look, it won't be edited, it won't annotated with English because it's just some giant basket of numbers. But it's tracking along with the inner monologue and saying like well, it thinks to itself, am I being watched? And in fact it normally concludes it is being watched on the training distribution, right? You have this whole distribution and so the explanation is saying it knows it's being watched and therefore it decides not stab anyone. So ideally that's what the explanation does. Again, not annotated English but just like a giant bag of numbers it's effectively talking about that process. So then when you have a new input on which the model says am I on train distribution? Like no, I'm not being watched. Then if you compare that against the explanation, the explanation said, like, hey, the reason it doesn't stab anyone is because it just constantly thinks that it's being trained. Like, this thing is always this nonlinear feature. The representation of am I being trained? Is constantly saying, yes, that's the explanation for why it never stabs people on the train distribution. And then on this test time, when you run it on the new input, it's like, does I think I'm on the train distribution. It says, no. You compare that against your explanation, actually, the explanation breaks down completely. Now, the explanation said it was precisely because it thought it was being trained, that it doesn't do anything really dangerous or doesn't stab anyone. And so once we see that single variable switch, we know that this is not just

anomaly, this is like, the anomaly that matters or is like, anomaly that is very relevant to the explanation.

Dwarkesh Patel
So this would require it to have this what is that cone that dogs have? But basically it would require this to have that MRI on it all the time. When it's deployed on every single call, you need to be running this loop of like, is this explanation still holding?

Paul Christiano
Yeah. For this story of flagging anomaly, you have to actually check at a sufficiently high frequency. So you can imagine checking all the time, or you could imagine saying, like, hey, actually, you need a lot of AI systems cooperating to not blow the whistle on each other or whatever. E g. You might have checked, like, hey, in training, every time something suspicious happens, the model flags it. Why does that happen? You have your explanation. You look at a new case, you could say, is it not flagging it for the normal reason, or is it a weird reason? So you could hope that even if a small fraction, even if you're running a small fraction of cases, if something bad is happening at scale, you can detect that. If you imagine, like, one in every thousand decisions, you check if there's, like, a robot rebellion going on or whatever, you get to look in on one in every thousand decisions. You see that something is really weird with that decision. And even if you didn't check many of them, you're like, okay, we need to back off.

Dwarkesh Patel
What happens if the AI, I mean, WebText, that includes your papers or maybe just, like, deduces the first principle that this is an alignment technique one might try, and then it says, okay, well, I'm going to engineer my activations. In a way, I'm trying to imagine how we might do this, but basically, even in deployment, I'm going to make my thought process so subtle that it doesn't raise any alarms. Is that a possibility?

Paul Christiano
Yeah. So I think the hope is to basically be robust or like, you know, when we're doing theoretical research, we are in a kind of worst Casey regime.

Dwarkesh Patel
So.

Paul Christiano
We are normally imagining that everything conspires insofar as it can. So insofar as could a brain be like doing some crazy thing where it tries to structure itself to look weird? That's totally in scope for ruling something out. And I guess the hope there is it's actually not that easy to tell a story. Like if you're like why did you not stab people on training time? Why are you not stabbing? Is that applying now? It's kind of a mind doesn't get to arbitrarily change

the actual reasons. It does something. Like there was some earlier step, maybe you could think of it as like at the first step where it's like now I'm going to try and do the sneaky thing to make my thoughts look this way. The actual explanation, like the actual causal explanation for the behavior is like there if you tried to give this reason it would have to maybe we should talk about the example in more detail to try and imagine how a brain would try and do this or how a mind would try and do this. But I think it is like the.

Dwarkesh Patel
Let's say it has this plan. So in training it keeps the part of it activates. A part of the.

Paul Christiano
Problem is if it ever does a check that's like am I in training or not, then your explanation for the behavior is going to have to factor through that check. And so if you ever do the check, the check becomes like an essential part of the explanation and then when the check fails, your explanation breaks down. So you've already lost the game if you did such a check.

Dwarkesh Patel
Is there a way it could make it seem like because if on every different input it has different activations, then can it make it seem like the different activations were a result of just a different input and not because some internal check went off?

Paul Christiano
Yeah, again, the hope is it needs to not only have them be like random different inputs, it needs to produce inputs with certain properties. Like it needs to deliberately create inputs that have this effect in the world, this intended effect in the world.

Dwarkesh Patel
So it basically creates a jailbreak for itself in some sense.

Paul Christiano
Yeah. Again the question is like it has to at test time be somehow producing activations that have the takeover effect and at training time be producing activations that don't look that takeovery or whatever. And so somehow there was a difference between those. It has to actually really do the work of creating that difference somewhere.

Dwarkesh Patel
Can you make more concrete for us what an explanation would look like physically? What does the explanation look like? Is it just a whole bunch of numbers itself? Is it a language? Is it code? What does the explanation look like?

Paul Christiano
Yeah, I also want to maybe step back a tiny bit and clarify that. I think this project is kind of crazily ambitious and the main reason, the overwhelming reason I think you should expect it to break down or fail is just because we have all these desires we have all the things we want out of this notion of explanation. But that's an incredibly hard research project which has a reasonable chance of being impossible. So I'm happy to talk about what the implications are but I want to flag but condition on failing I think it's most likely because, just like the things we wanted were either incoherent or intractably difficult.

Dwarkesh Patel
But what are the odds you think you'll succeed?

Paul Christiano
I mean, it depends a little bit what you mean by succeed. But if you, say, get explanations that are great and accurately reflect reality and work for all of these applications that we're imagining or that we are optimistic about, like kind of the best case success, I don't know, like 1020 percent something. And then there's like a higher probability of various intermediate results that provide value or insight without being, like, the whole dream. But I think the probability of succeeding in the sense of realizing the whole dream is quite low. Yeah in terms of what explanations look like physically or like the most ambitious plan, the most optimistic plan is that you are searching for explanations in parallel with searching for neural networks. So you have a parameterization of your space of explanations which mirrors the parameterization of your space of neural networks. Or you should think of as kind of similar to what is a neural network? It's some simple architecture where you fill in a trillion numbers and that specifies how it behaves. So to you should expect an explanation to be like a pretty flexible general skeleton that's saying pretty flexible general skeleton which just has a bunch of numbers you fill in. And what you are doing to produce an explanation is primarily just filling in these floating point numbers.

Dwarkesh Patel
When we conventionally think of explanations if you think of the explanation for why the universe moves this way it wouldn't be something that you could discover on some smooth evolutionary surface where you can climb up the hill towards the laws of physics. These are the laws of physics. You kind of just derive them from reverse principles. But in this case it's not like just a bunch of correlations between the orbits of different planets or something. Maybe the word explanation has a different I didn't even ask the question but maybe you can just speak to that.

Paul Christiano
Yeah, I think I basically sympathize. This is like there's some intuitive objections like, look, the space of explanations is this rigid, logical a lot of explanations have this rigid, logical structure where they're really precise and simple things govern complicated systems and

nearby simple things just don't work, and so on. And a bunch of things which feel totally different from this kind of nice, continuously parameterized space. And you can imagine interpretability on simple models where you're just like by gradient descent, finding feature directions that have desirable properties. But then when you imagine like, hey, now, that's like a human brain you're dealing with, that's like thinking logically about things. The explanation of why that works isn't going to be just like here with some featured directions. That's how I understood the basic confusion, which I share or sympathize with at least. So I think the most important high level point is I think basically the same objection applies to being like, how is GPT-4 going to learn to reason logically about something? You're like, well, look, logical reasoning that's like it's got rigid structure, it's doing ands and ors when it's called for, even though it just somehow optimized over this continuous space. And the difficulty or the hope is that the difficulty of these two problems are kind of like matched. So that is it's very hard to find these logicalish explanations because it's not a space that's easy to search over. But there are ways to do it. There's ways to embed discrete, complicated, rigid things in these nice, squishy continuous spaces that you search over. And in fact, to the extent that neural nets are able to learn the rigid logical stuff at all, they learn it in the same way. That is, maybe they're hideously inefficient, or maybe it's possible to embed this discrete reasoning in the space in a way that's not too inefficient, but you really want the two search problems to be of similar difficulty. And that's like the key hope overall. I mean, this is always going to be the key hope. The question is, is it easier to learn a neural network or to find the explanation for why the neural network works? I think people have the strong intuition that it's easier to find the neural network than the explanation of why it works. And that is really the I think we or at least exploring the hypothesis or interested in hypothesis that maybe those problems are actually more matched in difficulty.

Dwarkesh Patel
And why might that be the case?

Paul Christiano
This is pretty conjectural and complicated to express some intuitions. Maybe one thing is, I think a lot of this intuition does come from cases like machine learning. So if you ask about writing code and you're like, how hard is it to find code versus find the explanation the code is correct. In those cases, there's actually just like, not that much of a gap. Like the way a human writes a code is basically the same difficulty as find the explanation for why it's correct. In the case of ML, I think we just mostly don't have empirical evidence about how hard it is to find explanations of this particular type about why models work. We have a sense that it's really hard, but that's because we have this incredible mismatch where gradient descent is spending an incredible amount of compute searching for a model. And then some human is like looking at activate, looking at neurons or even some neural net is looking at neurons just like you have an incredible basically because you cannot define what an explanation is. You're not applying gradient descent to the search for explanations. So I think the ML case just actually shouldn't make you feel that pessimistic about the difficulty

of finding explanations. The reason it's difficult right now is precisely because you don't have any kind of you're not doing an analogous search process to find this explanation as you do to find the model. That's just like a first part of the intuition. Like when humans are actually doing design. I think there's not such a huge gap when in the ML case I think there is a huge gap. But I think largely for other reasons. A thing I also want to stress is that we just are open to there being a lot of facts that don't have particularly compact explanations. So another thing is when we think of finding an explanation in some sense we're setting our sites really low here. So if a human designed a random widget and was like, this widget appears to work well or if you search for a configuration that happens to fit into this spot really well it's like a shape that happens to mesh with another shape. You might be like, what's the explanation for why those things mesh? And we're very open to just being like that doesn't need an explanation. You just compute. You check that the shapes mesh and you did a billion operations and you check this thing worked. Or you're like, Why do these proteins? You're like, it's just because these shape like, this is a low energy configuration. And we're very open to in some cases, there's not very much more to say. So we're only trying to explain cases where kind of the surprise intuitively is very large. So, for example, if you have a neural net that gets a problem correct a neural net with a billion parameters that gets a problem correct on every input of length 1,000 in some sense, there has to be something that needs explanation there because there's, like, too many inputs for that to happen by chance alone. Whereas if you have a neural net that gets something right on average or gets something right in merely a billion cases, that actually can just happen by coincidence. GPT-4 can get billions of things right by coincidence because it just has so many parameters that are adjusted to fit the data.

Dwarkesh Patel
So a neural net that is initialized completely randomly the explanation for that would just be the neural net itself.

Paul Christiano
Well, it would depend on what behaviors it had. So we're always, like, talking about an explanation of some behavior from a model, right?

Dwarkesh Patel
And so it just has a whole bunch of random behaviors. So it'll just be like an exponentially large explanation relative to the weights of the model.

Paul Christiano
Yeah, I think there just aren't that many behaviors that demand explanation. Like most things a random neural net does are kind of what you'd expect from, like, a random if you treat it just like a random function, then there's nothing to be explained. There are some behaviors that demand explanation. But anyway, random neural net is pretty uninteresting. That's part of the hope is it's kind of easy to explain features of the random neural net.

Dwarkesh Patel
Okay, so that's interesting. So the smarter or more ordered the neural network is, the more compressed the explanation.

Paul Christiano
Well, it's more like the more interesting the behaviors to be explained. So the random neural net just doesn't have very many interesting behaviors that demand explanation. And as you get smarter, you start having behaviors that are like, you start having some correlation with the simple thing and then that demands explanation. Or you start having some regularity in your outputs, and that demands explanation. So these properties kind of emerge gradually over the course of training that demand explanation. I also, again, want to emphasize here that when we're talking about searching for explanations, this is some dream. We talk to ourselves. Like, why would this be really great if we succeeded? We have no idea about the empirics on any of this. So these are all just words that we think to ourselves and sometimes talk about to understand. Would it be useful to find a notion of explanation? And what properties would we like this notion of explanation to have? But this is really like, speculation and being out on a limb almost all of our time, day to day is just thinking about cases much, much simpler even than small neural nets or thinking about very simple cases and saying, what is the correct notion? What is the right heuristic estimate in this case? Or how do you reconcile these two apparently conflicting explanations?

Dwarkesh Patel
Is there a hope that if you have a different way to make proofs now that you can actually have heuristic arguments where instead of having to prove the Riemann hypothesis or something you can come up with a probability of it in a way that is compelling and you can publish? So would it just be a new way to do mathematics? A completely new way to prove things in mathematics?

Paul Christiano
I think most claims in mathematics that mathematicians believe to be true already have fairly compelling heuristic arguments like the Riemann hypothesis. It's actually just there's kind of a very simple argument that the Riemann hypothesis should be true unless something surprising happens. And so a lot of math is about saying, like, okay, we did a little bit of work to find the first pass explanation of why this thing should be true. And then, for example, in the case of the Riemann hypothesis, the question is, do you have this weird periodic structure in the primes? And you're like, well, look, if the primes were kind of random you obviously wouldn't have any structure like that. Just how would that happen? And then you're like, well, maybe there's something and then the whole activity is about searching for can we rule out anything? Can we rule out any kind of conspiracy that would break this result? So I think the mathematicians just wouldn't be very surprised or wouldn't care that much. And this is related to the motivation for the project. I think just in a lot of

domains, in a particular domain, people already have norms of reasoning that work pretty well and match roughly how we think these heuristic arguments should work.

Dwarkesh Patel
But it would be good to have more concrete sense, like if you could say instead of, well, we think RSA is fine, to being able to say, here's the probability that RSA is fine.

Paul Christiano
Yeah. My guess is these will not. Like, the estimates you get out of this would be much, much worse than the estimates you'd get out of just normal empirical or scientific reasoning where you're using a reference class and saying, how often do people find algorithms for hard? Like, I think what this argument will give you for is RSA fine? Is going to be like, well, RSA is fine. Unless it isn't. Unless there's some additional structure in the problem that an algorithm can exploit, then there's no algorithm. But very often the way these arguments work, so for neural nets as well, is you say, like, look, here's an estimate about the behavior, and that estimate is right unless there's another consideration we've missed. And the thing that makes them so much easier than proofs is just say, like, here's a best guess, given what we've noticed so far, but that best guess can be easily upset by new information. And that's both what makes them easier than proofs, but also what means they're just, like, way less useful than proofs for most cases. I think neural nets are kind of unusual in being a domain where we really do want to do systematic, formal reasoning, even though we're not trying to get a lot of confidence, we're just trying to understand even roughly what's going on.

Dwarkesh Patel
But the reason this works for alignment but isn't that interesting for the Riemann hypothesis, where if in the RSA case, you say, well, the RSA is fine unless the estimate is wrong, it's like, well, okay, well, it would tell us something new. But in the alignment case, if the estimate is, this is what the output should be, unless there's some behavior I don't understand, you want to know? In the case, unless there's some behavior you don't understand that's not like, oh, whatever. That's the case in which it's not aligned.

Paul Christiano
Yeah, I mean, maybe one way of putting it is just like, we can wait until we see this input, or like, you can wait until you see a weird input and say, okay, weird input, do something we didn't understand. And for our say, that would just be a trivial test. You're just like, in some cases algorithms would be like is it a thing? Whereas for neural net in some cases it is either very expensive to tell or it's like you actually don't have any other way to tell. Like you checked in easy cases and now you're on a hard case so you don't have a way to tell if something has gone wrong. Also, I would clarify that I think it is interesting for the Riemann hypothesis I would say the current state, particularly in number theory, but maybe in quite a lot of math, is like there are informal heuristic arguments for pretty much all the open

questions people work on but those arguments are completely informal. So that is like I think it's not the case that there's like here's the norms of informal reasoning or the norms of heuristic reasoning and then we have arguments that a heuristic argument verifier could accept. It's just like people wrote some words. I think those words like my guess would be like 90 of the things mathematicians accept as really compelling filling heuristic arguments are correct and if you actually formalize them you'd be like some of these aren't quite right, or here's some corrections or here's which of two conflicting arguments is right? I think there's something to be learned from it. I don't think it would be like mind blowing. No.

Dwarkesh Patel
When you have it completed, how big would this heuristic estimator the rules for this heuristic estimator mean, I know like when Russell and who was the other guy when they did the rules? Yeah, wasn't it like literally they had like a bucket or a wheelbarrow with all the papers.

Paul Christiano
But how big would I mean, mathematical foundations are quite simple in the end. At the end of the day it's like how many symbols? I don't know, it's hundreds of symbols or something that go into the entire foundations and the entire rules of reasoning for like there's a sort of built on top of first order logic but the rules of reasoning for first order logic are just like another hundreds of symbols or 100 lines of code or whatever. I'd say I have no idea. We are certainly aiming at things that are just not that complicated and my guess is that the algorithms we're looking for are not that complicated. Most of the complexity is pushed into arguments not in this verifier or estimator.

Dwarkesh Patel
So for this to work you need to come up with an estimator which is a way to integrate different heuristic arguments together.

Paul Christiano
Has to be a machine that takes its input. Like first it takes an input argument, decides what it believes in light of it, which is kind of like saying was it compelling? But second, it needs to take 4 of those and then say here's what I believe in light of all four, even though there's a different estimation strategies that produce different numbers and that's like a lot of our life is saying like well, here's a simple thing that seems reasonable. And here's a simple thing that seems reasonable. What are you doing? There's supposed to be a simple thing that unifies them both. And the obstruction to getting that is understanding what happens when these principles are slightly intention and how do we deal?

Dwarkesh Patel
Yeah, that seems super interesting. We'll see what other applications it has. I don't know, like computer security and code checking. If you can actually say this is how safe we think a code is.

Paul Christiano
In a very formal way, my guess is we're not going to add I mean, this is both a blessing and a curse. It's a curse. And you're like, well, that's sad. Your thing is not that useful, but a blessing and not useful things are easier. My guess is we're not going to add that much value in most of these domains. Most of the difficulty comes from a lot of code that you'd want to verify. Not all of it, but a significant part. It's just like the difficulty of formalizing the proof is like the hard part and actually getting all of that to go through and we're not going to help even the tiniest bit with that, I think. So this would be more helpful if you have code that uses simulations, you want to verify some property of a controller that involves some numerical error or whatever you need to control the effects of that error. That's where you start saying like, well, heuristically, if the errors are independent - blah, blah, blah.

Dwarkesh Patel
Yeah, you're too honest to be a salesman, Paul.

Paul Christiano
This is kind of like sales to us, right? If you talk about this idea, people are like, why would that not be the coolest thing ever and therefore impossible? And we're like, well, actually it's kind of lame and we're just trying to pitch it's way lamer than it sounds. And that's really important to why it's possible, is being like, it's really not going to blow that many people's. I mean, I think it will be cool. I think it will be like very if we succeed will be very solid, like metamathematics or theoretical computer science or whatever. But I don't think I think the mathematicians already do this reasoning and they mostly just love proofs. I think the physicists do a lot of this reasoning, but they don't care about formalizing anything. I think in practice, other difficulties are almost always going to be more salient. I think this is of most interest by far for interpretability and ML and I think other people should care about it and probably will care about it if successful. But I don't think it's going to be the biggest thing ever in any field or even that huge a thing. I think this would be a terrible career move given the ratio of difficulty to impact. I think theoretical computer science, it's probably a fine move. I think in other domains it just wouldn't be worth we're going to be working on this for years, at least in the best case.

Dwarkesh Patel
I'm laughing because my next question was going to be like a set up for you to explain if this grad student wants to work on this.

Paul Christiano
I think theoretical computer science is an exception where I think this is like, in some sense, like what the best of theoretical computer science is like. So you have all this reason you have this because it's useless. Like an analogy. I think one of the most successful sagas in theoretical computer science is like formalizing the notion of an interactive proof system. And it's like you have some kind of informal thing that's interesting to understand, and you want to pin down what it is and construct some examples and see what's possible and what's impossible. And this is like I think this kind of thing is the bread and butter of the best parts of theoretical computer science. And then again, I think mathematicians it may be a career mistake because the mathematicians only care about proofs or whatever, but that's a mistake in some sense. Aesthetically, it's successful. I do think looking back and again, part of why it's a mistake is such a high probability we wouldn't be successful. But I think looking back, people would be like, that was pretty cool, although not that cool. Or we understand why it didn't happen given the epistemic, like what people cared about in the field, but it's pretty cool now.

Dwarkesh Patel
But isn't it also the case that didn't Hardy write in that all this prime shit is both not useless, but it's fun to do, and it turned out that all the cryptography is based on all that prime shit. So I don't know. But anyways, I'm trying to set you up so that you can tell and forget about if it doesn't have applications in all those other fields. It matters a lot for Alignment and that's why I'm trying to set you up to talk about if I think a lot of smart people listen to this podcast. If they're a math or CS grad student and has gotten interested in this. Are you looking to potentially find talent to help you with this? Yeah, maybe we'll start there. And then I also want to ask you if I think also maybe people who can provide funding might be listening to the podcast. So to both of them, what is your pitch?

Paul Christiano
We're definitely hiring and searching for collaborators. I think the most useful profile is probably a combination of intellectually interested in this particular project and motivated enough by alignment to work on this project, even if it's really hard. I think there are a lot of good problems. The basic fact that makes this problem unappealing to work on I'm a really good salesman, but whatever. I think the only reason this isn't a slam dunk thing to work on is that there are not great examples. So we've been working on it for a while, but we do not have beautiful results as of the recording of this podcast. Hopefully by the time it airs, you completely script. They've had great results since then, but.

Dwarkesh Patel
It was too long to put in the margins of the podcast.

Paul Christiano

Yeah, with luck. Yeah. So I think it's hard to work on because it's not clear what a success looks like. It's not clear if success is possible. But I do think there's a lot of questions. We have a lot of questions and I think the basic setting of, like, look, there are all of these arguments. So in mathematics, in physics, in computer science are just a lot of examples of informal heuristic arguments. They have enough structural similarity that it looks very possible that there is like a unifying framework, that these are instances of some general framework and not just a bunch of random things. Like not just a bunch of it's not like so, for example, for the prime numbers, people reason about the prime numbers as if they were like a random set of numbers. One view is like, that's just a special fact about the primes, they're kind of random. A different view is like, actually it's pretty reasonable to reason about an object as if it was a random object as a starting point. And then as you notice structure, like revised from that initial guess and it looks like to me, the second perspective is probably more right. It's just like reasonable to start off treating an object as random and then notice perturbations from random. Like, notice structure the object possesses and the primes are unusual and that they have fairly little additive structure. I think it's a very natural theoretical project. There's like a bunch of activity that people do. It seems like there's a reasonable chance there's something nice to say about unifying all of that activity. I think it's a pretty exciting project. The basic strike against it is that it seems really hard. Like if you were someone's advisor, I think you'd be like, what are you going to prove if you work on this for the next two years? And they'd be like, there's a good chance. Nothing. And then it's not what you do if you're a PhD student. Normally you aim for those high probabilities of getting something within a couple of years. The flip side is it does feel I mean, I think there are a lot of questions. I think some of them we're probably going to make progress on. So I think the pitch is mostly like, are some people excited to get in now? Or are people more like, let's wait to see. Once we have one or two good successes to see what the pattern is and become more confident, we can turn the crank to make more progress in this direction. But for people who are excited about working on stuff with reasonably high probabilities of failure and not. Really understanding exactly what you're supposed to do. I think it's a pretty good project. I feel like if people look back if we succeed and people are looking back in 50 years on what was the coolest stuff happening in math or theoretical computer science, there will be, like a reasonable this will definitely be, like, in contention. And I would guess for lots of people would just seem like the coolest thing from this period of a couple of years or whatever.

Dwarkesh Patel

Right. Because this is a new method in so many different fields from the ones you met physics, math, theoretical computer science, I don't know because what is the average math PhD working on? Right? He's working on a subset of a subset of something I can't even understand or pronounce. But math is quite esoteric. But yeah, this seems like, I don't know, even small chance of it working. You shouldn't forget about the value for alignment. But even without that, this is such a cool if this works, it's like a really big deal.

Paul Christiano
There's a good chance that if I had my current set of views about this problem and didn't care about alignment and had the career safety to just spend a couple of years thinking about it, spend half my time for like five years or whatever, that I would just do that. I mean, even without caring at all about alignment, it's a very nice problem. It's very nice to have this library of things that succeed where they feel so tantalizingly close to being formalizable, at least to me, and such a natural setting, and then just have so little purchase on it. There aren't that many really exciting feeling frontiers in theoretical computer science.

Dwarkesh Patel
And then smart person doesn't have to be a grasshood, but a smart person is interested in this. What should they do? Should they try to attack some open problem you have put on your blog? Or should it what is the next step?

Paul Christiano
Yeah, I think a first path step. There's different levels of ambition or whatever, different ways of approaching a problem. But we have this write up from last year or I guess eleven months ago or whatever on formalizing, the presumption of independence that provides, like, here's kind of a communication of what we're looking for in this object. And I think the motivating problem is saying here's a notion of what an estimator is and here's what it would mean for an estimator to capture some set of informal arguments. And a very natural problem is just try and do that. Go for the whole thing, try and understand and then come up with hopefully a different approach or then end up having context from a different angle on the kind of approach we're taking. I think that's a reasonable thing to do. I do think we also have a bunch of open problems, so maybe we should put up more of those open problems. I mean, the main concern with doing so is that for any given one, we're like, this is probably hopeless. Like, put up a prize earlier in the year for an open problem, which tragically, I mean, I guess the time is now to post the debrief from that, or I owe it from this weekend. I was supposed to do that, so I'll probably do it tomorrow, but no one solved it. It's sad putting out problems that are hard or like I don't we could put out a bunch of problems that we think might be really hard.

Dwarkesh Patel
But what was that famous case of that statistician who it was like, some PhD student who showed up late to a class and he saw some problems on the board and he thought they were homework, and then they were actually just open problems, and then he solved them because he thought they were homework.

Paul Christiano
Right, yeah. I mean, we have much less information that these problems are hard. Again, I expect the solution to most of our problems to not be that complicated. And we've been working on it in some sense for a really long time. Total years of full-time equivalent work

across the whole team is like probably like 3 years of full-time equivalent work in this area spread across a couple of people. But that's very little compared to a problem. It is very easy to have a problem where you put in 3 years of full-time equivalent work. But in fact, there's still an approach that's going to work quite easily with like, 3 to six months if you come at a new angle. And we've learned a fair amount from that that we could share, and we probably will be sharing more over the coming months.

Dwarkesh Patel
As far as funding goes, is this something where, I don't know, if somebody gave you a whole bunch of money that would help? Or does it not matter how many people are working on this, by the way?

Paul Christiano
So, we have been - right now, there's 4 of us full-time and we're hiring for more people.

Dwarkesh Patel
And then is funding that would matter?

Paul Christiano
I mean, funding is always good. We're not super funding constrained right now. The main effect of funding is it will cause me to continuously and perhaps indefinitely delay fundraising. Periodically. I'll set out to be interested in fundraising and someone will be like, offer a grant, and then I will get to delay for another six months or fundraising or nine months, or you can you can delay the time at which Paul needs to think for some time about fundraising.

Dwarkesh Patel
Well, one question I think would be interesting to ask, you know, I think people can talk vaguely about the value of theoretical research and how it contributes to real world applications and you can look at historical examples or something, but you are somebody who actually has done this in a big way. Like RLHF is something you developed and then it actually has got into an application that has been used by millions of people. Tell me about just that pipeline. How can you reliably identify theoretical problems that will matter for real world applications? Because it's one thing to read about touring or something and the Halting problem, but here you'd have the real thing.

Paul Christiano
Yeah, I mean, it is definitely exciting to have worked on a thing that has a real world impact. The main caveat I'd provide is, like, RLHF is very simple compared to many things. And so the motivation for working on that problem was, like, look, this is how it probably should work, or this is a step in some progression. It's unclear if it's, like, the final step or something, but it's a very natural thing to do that people probably should be and probably

will be doing. I'm saying, if you want to talk about crazy stuff, it's good to help make those steps happen faster, and it's good to learn about. There's lots of issues that occur in practice, even for things that seem very simple on paper, but mostly, like, the story of it's just like, yeah, I think my sense of the world is things that look like good ideas on paper, just, like, often are harder than they look. But the world isn't that far from what makes sense on paper. Like, large language models look really good on paper, and RLH looks really good on paper. And these things, I think, just work out in a way that's yeah, I think people maybe overestimate or, like, maybe it's kind of a trope, but people talk about, like, it's easy to underestimate how much gap there is to practice, like, how many things will come up that don't come up in theory. But it's also easy to overestimate how inscrutable the world is. Like, the things that happen mostly are things that do just kind of make sense. Yeah, I feel like most ML implementation does just come down to a bunch of detail, though, of, like, build a very simple version of the system, understand what goes wrong, fix the things that go wrong, scale it up, understand what goes wrong. And I'm glad I have some experience doing that, but I think that does cause me to be better informed about what makes sense in ML and what can actually work. But I don't think it caused me to have a whole lot of deep expertise or deep wisdom about how to close the gap.

Dwarkesh Patel
Yeah, but is there some tip on identifying things like RLHF which actually do matter, versus making sure you don't get stuck in some theoretical problem that doesn't matter? Or is it just coincidence? Or I mean, is there something you can do in advance to make sure that the thing is useful?

Paul Christiano
I don't know if the RLHF story is, like, the best success case or something, but because the capabilities maybe I'd say more profoundly, like, again, it's just not that hard a case. It's a little bit unfair to be like, I'm going to predict the thing, which I pretty much think it was going to happen at some point. And so it was mostly a case of acceleration, whereas the work we're doing right now is specifically focused on something that's kind of crazy enough that it might not happen. Even if it's a really good idea or challenging enough, it might not happen. But I'd say in general, and this draws a little bit on more broad experience more broadly in theory, it's just like a lot of the times when theory fails to connect with practice. It's just kind of clear it's not going to connect. If you like, try if you actually think about it and you're like, what are the key constraints in practice? Is theoretical problem we're working on actually connected to those constraints? Is there something that is possible in theory that would actually address real world issues? I think the vast majority as a theoretical computer scientist, the vast majority of theoretical computer science has very little chance of ever affecting practice. But also it is completely clear in theory that has very little chance of affecting practice. Most of theory fails to affect practice, not because of all the stuff you don't think of, but just because you could call it like dead on arrival, but you could also be like, it's not really the point. It's just like mathematicians also are like, they're not trying to

affect practice and they're not like, why does my number theory not affect practice? It was kind of obvious. I think the biggest thing is just like, actually caring about that and then learning at least what's basically going on in the actual systems you care about and what are actually the important constraints. And is this a real theoretical problem? The basic reason most theory doesn't do that is just like, that's not where the easy theoretical problems are. So I think theory is instead motivated by like, we're going to build up the edifice of theory and sometimes there'll be Opportunistic. Opportunistically we'll find a case that comes close to practice, or we'll find something practitioners are already doing and try and bring into our framework or something. But theory of change is mostly not this thing is going to make into practice. It's mostly this is going to contribute to the body of knowledge that will slowly grow. And sometimes opportunistically yields important results.

Dwarkesh Patel
How big do you think a seed AI would be? What is the minimum sort of encoding of something that is as smart as a human?

Paul Christiano
I think it depends a lot what substrate it gets to run on. So if you tell me how much computation does it get before or what kind of real world infrastructure does it get? You could ask what's the shortest program, which if you run it on a million h 100s connected in a nice network with a hospitable environment will eventually go to the stars. But that seems like it's probably on the order of tens of thousands of bytes or I don't know if I had to guess the median, I'd guess 10,000 bytes.

Dwarkesh Patel
Wait, the specification or the compression of just the program?

Paul Christiano
A program which went wrong. Oh, got it. But that's going to be like, really cheatsy. So they ask, what's the thing that has values and will expand and roughly preserve its value? Because that thing, the 10,000 byte thing, will just lean heavily on evolution and natural selection to get there for that. Like, I don't know, million bytes, million bytes, 100,000 bytes, something like that.

Dwarkesh Patel
Do you think AI lie detectors will work where you kind of just look at the activations and not find explanations in the way you were talking about with heuristics, but literally just like, here's what truth looks like, here's what lies look like. Let's just segregate the lane space and see if we can identify the two.

Paul Christiano
Yeah, I think to separate the like just train a classifier to do it is a little bit complicated for a few reasons and may not work. But if you just brought them to space and say like, hey, it's like you want to know if someone's lying, you get to interrogate them, but also you get to rewind them arbitrarily and make a million copies of them. I do think it's pretty hard to lie successfully. You get to look at their brain even if you don't quite understand what's happening. You get to rewind them a million times. You get to run all those parallel copies into gradient descent or whatever. I think there's a pretty good chance that you can just tell if someone is lying, like a brain emulation or an AI or whatever, unless they were aggressively selected. If it's just they are trying to lie well rather than it's like they were selected over many generations to be excellent at lying or something, then your ML system hopefully didn't train it a bunch to lie. And you want to be careful about whether your training scheme effectively does that. Yeah, that seems like it's more likely than not to succeed.

Dwarkesh Patel
And how possible do you think it will be for us to specify human verifiable rules for reasoning such that even if the AI is super intelligent, we can't really understand why it does certain things. We know that the way in which it arises at these conclusions is valid. Like, if it's trying to persuade us to something, we can be like, I don't understand all the steps, but I know that this is something that's valid and you're not just making shit up.

Paul Christiano
That seems very hard if you wanted to be competitive with learned reasoning, it depends a little bit exactly how you set it up. But for the ambitious versions of that, let's say it would address the alignment problem, they seem pretty unlikely, like 5% kind of thing.

Dwarkesh Patel
Is there an upper bound on intelligence? Not in the near term, but just like super intelligence at some point. How far do you think that can go?

Paul Christiano
It seems like it's going to depend a little bit on what is meant by intelligence. It kind of reads as a question that's similar to is there an upper bound on strength or something? There are a lot of forms. I think it's like the case that I think there are sort of arbitrarily smart input output functionalities and then if you hold fixed the amount of compute, there is some smartest one if you're just like, what's the best set of ten to the 40th operations? There's only finitely many of them. So some best one for any particular notion of best that you have in mind? So I guess I'm just like for the unbounded question where you're allowed to use arbitrary description complexity and compute, like probably no and for the I mean, there is some optimal conduct if you're like I have some goal in mind and I'm just like, what action best achieves it? If you imagine like a little box embedded in the universe, I think there is

kind of just like an optimal input output behavior. So I guess in that sense I think there is an upper bound, but it's not saturatable in the physical universe because it's definitely exponentially slow, right?

Dwarkesh Patel
Yeah. Because of comms or other things or heat. It just might be physically impossible to instagram something smarter than this.

Paul Christiano
Yeah, I mean, like, for example, if you imagine what the best thing is, it would almost certainly involve just like simulating every possible universe. It might be in modular moral constraints, which I don't know if you want to include like so that would be very slow. It would involve simulating like, I don't know exactly how slow, but like double exponential very slow.

Dwarkesh Patel
Carl Schulman laid out his picture of the intelligence explosion in the seven hour episode. I know you guys have talked a lot. What about his basic is? Do you have some main disagreements? Is there some crux that you guys have explored?

Paul Christiano
It's related to our timelines discussion from yeah, I think the biggest issue is probably error bars where Carl has a very software focused, very fast kind of takeoff picture. And I think that is plausible, but not that likely. I think there's a couple of ways you could perturb the situation and my guess is one of them applies. So maybe I have like I don't know exactly what Carl's probability is. I feel like Carl's going to have like a 60% chance on some crazy thing that I'm only going to assign like a 20% chance to or 30% chance or something. And I think those kinds of perturbations are like one, how long a period is there of complementarity between AI capabilities and human capabilities which will tend to soften takeoff? Two, how much diminishing returns are there on software progress, such that is a broader takeoff involving scaling, electricity production and hardware production. Is that likely to happen during takeoff, where I'm more like 50-50 or more stuff like this?

Dwarkesh Patel
Yeah. Okay, so is it that you think the alternate constraints will be more hard? The basic case he's laid out is that you can just have a sequence of things like flash attention or Moe, and you can just keep stacking these kinds of things on.

Paul Christiano
I'm very unsure if you can keep stacking them or like it's kind of a question of what's like, the returns curve and Carl has some inference from historical data or some way he'd extrapolate the trend. I am more like 50-50 on whether the software only intelligence

explosion is even possible, and then like a somewhat higher probability that it's slower than why.

Dwarkesh Patel
Do you think it might not be possible?

Paul Christiano
Well, the entire question is like: If you double R&D effort, do you get enough additional improvement to further double the efficiency? And that question will itself be a function of your hardware base, like how much hardware you have. And the question is like, at the amount of hardware we're going to have and the level of sophistication we have as the process begins. Is it the case that each doubling of actually the initial only depends on the hardware, or like, each level of hardware will have some place at this dynamic asymptotes so the question is just like, for how long? Is it the case that each doubling of R&D at least doubles the effective output of your AI research population? And I think I have a higher probability on that. I think it's kind of close. If you look at the Empirics, I think the Empirics benefit a lot from continuing hardware scale up so that the effective R&D stock is significantly smaller than it looks, if that makes sense.

Dwarkesh Patel
What are the Empirics you're referring to?

Paul Christiano
So there's kind of two sources of evidence. One is like, looking across a bunch of industries at like, what is the general improvement with each doubling of either R&D investment or experience, where it is quite exceptional to have a field with not anyway. It's pretty good to have a field where each time you double R&D investment, you get a doubling of efficiency. The second source of evidence is on actual algorithmic improvement in ML, which is obviously much, much scarcer. And there you can make a case that it's been like each doubling of R&D has given you roughly a forex or something increase in computational efficiency. But there's a question of how much that benefits. When I say the effect of R&D stock is smaller, I mean we scale up. You're doing a new task like every couple years, you're doing a new task because you're operating a scale much larger than the previous scale. And so a lot of your effort is how to make use of the new scale. So if you're not increasing your installed hardware base or just flat at a level of hardware, I think you get much faster diminishing returns than people have gotten historically. I think Carl agrees, in principle, this is true. And then once you make that adjustment, I think it's, like, very unclear where the empirics shake out. I think Carl has thought about these more than I am, so I should maybe defer more. But anyway, I'm at like 50-50 on that.

Dwarkesh Patel
How have your timelines changed over the last 20 years?

Paul Christiano
Last 20 years?

Dwarkesh Patel
Yeah. How long have you been working on anything related to AI?

Paul Christiano
So I started thinking about this stuff in 2010 or so. So I think my earliest timeline prediction will be in 2011. I think in 2011, my rough picture was like, we will not have insane AI in the next ten years. And then I get increasingly uncertain after that. But we converged to 1% per year or something like that. And then probably in 2016, my take was, like, we won't have crazy AI in the next five years, but then we converged to, like, one or 2% per year after that. Then in 2019, I guess I made a round of forecasts where I gave like 30% or something to 25% to crazy Eye by 2040 and like 10% by 2030 or something like that. So I think my 2030 probability has been kind of stable, and my 2040 probability has been going up. And I would guess it's too sticky. I guess that 40% I gave at the beginning is just, like, from not having updated recently enough, and I maybe just need to sit down. I would guess that should be even higher. I think, like 15% in 2030. I'm not feeling that bad about this is just like, each passing year is, like, a big update against 2030. We don't have that many years left, and that's roughly counterbalanced with AI going pretty well. Whereas for the 2040 thing, the passing years are not that big a deal. And as we see that things are basically working, that's like, cutting out a lot of the probability of not having AI by 2040. My 2030 probability up a little bit, like, maybe twice as high as it used to be or something like that. My 2040 probability up much more significantly.

Dwarkesh Patel
How fast do you think we can keep building fabs to keep up with the eye demand?

Paul Christiano
Yeah, I don't know much about any of the relevant areas. My best guess is my understanding is right now, like 5% or something of the next year's total or best process. Fabs will be making AI hardware, of which only a small fraction will be going into very large training runs. Like, only a couple. So maybe a couple of percent of total output, and then that represents maybe like 1% of total possible output. A couple of percent of leading process 1% of total or something. I don't know if that's right, but I think it's like the rough ballpark we're in. I think things will be pretty fast. You scale up for the next order of magnitude or two from there because you're basically just shifting over other stuff. My sense is it would be like years of delay. There's like, multiple reasons that you expect years of delay for going past that, maybe even at that you start having. Yeah, there's just a lot of problems. Like building new fabs is quite slow and I don't think there's like, TSMC is not like, planning on increases in total demand driven by AI. Like kind of conspicuously not planning on it. I don't think anyone else is really ramping up production in anticipation think and then similarly just building

data centers of that size seems like very, very hard and also probably has multiple years of delay.

Dwarkesh Patel
What does your portfolio look like?

Paul Christiano
I've tried to get rid of most of the AI stuff that's plausibly implicated in policy work or like CEG advocacy on the RSP stuff for my involvement with Anthropic.

Dwarkesh Patel
What would it look like if you.

Paul Christiano
Had no conflicts of interest and no inside? Like, I also still have a bunch of hardware investments which I need to think about, but I don't know a lot of TSMC. I have a chunk of Nvidia, although I just keep betting against Nvidia constantly since 2016 or something. I've been destroyed on that bet. Although AMD has also done fine. The case now is even easier, but it's similar to the case in the old days, just a very expensive company. Given the total amount of R&D investment they've made, they have like, whatever, a trillion dollar valuation or something that's like very high. So the question is, how expensive is it to make a TPU? So it actually outcompetes H 100 or something. And I'm like, wow, it's real level, high level of incompetence if Google can't catch up fast enough to make that trillion dollar valuation not justified.

Dwarkesh Patel
Whereas with TSMC they have a harder remote, you think?

Paul Christiano
Yeah, I think it's a lot harder, especially if you're in this regime where you're trying to scale up. So if you're unable to build fabs, I think what will take a very long time to build as many fabs as people want, the effect of that will be to bid up the price of existing fabs and existing semiconductor manufacturing equipment. And so just those hard assets will become spectacularly valuable, as will the existing GPUs and the actual yeah, I think it's just hard. That seems like the hardest asset to scale up quickly. So it's like the asset, if you have like a rapid run up, it's the one that you'd expect to most benefit. Whereas Nvidia's stuff will ultimately be replaced by either better stuff made by humans or stuff made by with AI assistance. Like the gap will close even further as you build AI systems.

Dwarkesh Patel
Right. Unless Nvidia is using those systems.

Paul Christiano
Yeah, the point is just like anybody will so dwarf past R&D and there's like just not that much stickiness. There's less stickiness in the future than there has been in the yeah, I don't know. So I don't want to not commenting for any private information just in my gut having caveatted, this is like the single bet I've most okay not including Nvidia in that portfolio.

Dwarkesh Patel
And final question, there's a lot of schemes out there for alignment and I think just like a lot of general takes and a lot of this stuff is over my head where I think I literally it took me like weeks to understand the mechanistic anomaly stuff you work on without spending weeks.

Paul Christiano
How do you detect bullshit?

Dwarkesh Patel
People have explained their schemes to me and I'm like, honestly, I don't know if it makes sense or not with you. I'm just like I trust Paul enough that I think there's probably something here if I try to understand this enough. But how do you detect bullshit?

Paul Christiano
Yeah, so I think it's depends on the kind of work. So for the kind of stuff we're doing, my guess is like most people there's just not really a way you're going to tell whether it's bullshit. So I think it's important that we don't spend that much money on the people we want to hire are probably going to dig in in depth. I don't think there's a way you can tell whether it's bullshit without either spending a lot of effort or leaning on deference with empirical work. It's interesting in that you do have some signals of the quality of work. You can be like, does it work in practice? Does the story? I think the stories are just radically simpler and so you probably can evaluate those stories on their face. And then you mostly come down to these questions of like, what are the key difficulties? Yeah, I tend to be optimistic when people dismiss something because this doesn't deal with a key difficulty or this runs into the following insurable obstacle. I tend to be a little bit more skeptical about those arguments and tend to think, like, yeah, something can be bullshit because it's not addressing a real problem that's I think the easiest way this is a problem someone's interested in that's just not actually an important problem, and there's no story about why it's going to become an important problem. E g, like it's not a problem now and won't get worse or it is maybe a problem now, but it's clearly getting better. That's like one way and then conditioned on passing that bar, like dealing with something that actually engages with important parts of the argument for concern and then actually making sense empirically. So I think most work is anchored by source of feedback is like actually engaging with real models. So it's like, does it make sense how to engage with real models? And does the story about how it deals with key difficulties actually make sense? I'm pretty liberal past there. I think it's really hard to, like, eg. People look at mechanistic, interpretability and be like, well, this obviously can't

succeed. And I'm like, I don't know. How can you tell? It obviously can't succeed. I think it's reasonable to take total investment in the field. How fast is it making progress? How does that pencil I think most things people work on, though, actually pencil pretty fine. They look like they could be reasonable investments. Things are not, like, super out of whack.

Dwarkesh Patel
Okay, great. This is, I think, a good place to close. Paul, thank you so much for your time.

Paul Christiano
Yeah, thanks for having me. It was good chatting.

Dwarkesh Patel
Yeah, absolutely.