Dwarkesh Podcast #78 - Joe Carlsmith - Preparing for an AI civilization

Published - August 22, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I'm chatting with Joe Carlsmith. He's a philosopher and, in my opinion, a capital-G great philosopher. You can find his essays at joecarlsmith.com.

So we have GPT-4, and it doesn't seem like a paperclipper thing. It understands human values. In fact, you can have it explain why being a paperclipper is bad or ask it to explain why the galaxy shouldn't be turned into paperclips.

What has to happen such that eventually we have a system that takes over and converts the world into something valueless?

Joe Carlsmith
When I'm thinking about misaligned AIs—or the type that I'm worried about—I'm thinking about AIs with a relatively specific set of properties related to agency, planning, awareness, and understanding of the world.

One key aspect is the capacity to plan and make relatively sophisticated plans based on models of the world, where those plans are evaluated according to criteria. That planning capability needs to be driving the model's behavior. There are models that are, in some sense, capable of planning. But when they give output, it's not like that output was determined by some process of planning, like, “Here's what will happen if I give this output, and do I want that to happen?"

The model needs to really understand the world. It needs to really be like, “Okay, here’s what will happen. Here I am. Here’s the politics of the situation.” It needs to have this kind of situational awareness to evaluate the consequences of different plans.

Another thing to consider is the verbal behavior of these models. When I talk about a model's values, I'm referring to the criteria that end up determining which plans the model pursues. A model's verbal behavior—even if it has a planning process (which GPT-4, I think, doesn't in many cases)—doesn't necessarily reflect those criteria.

We know that we're going to be able to get models to say what we want to hear. That's the magic of gradient descent. Modulo some difficulties with capabilities, you can get a model to output the behavior that you want. If it doesn't, then you crank it until it does.

I think everyone admits that for suitably sophisticated models, they're going to have a very detailed understanding of human morality. The question is, what relationship is there between a model's verbal behavior—which you've essentially clamped, you're forcing the model to say certain things—and the criteria that end up influencing its choice between plans?

I'm pretty cautious about assuming that when it says the thing I forced it to say—or when gradient descent has shaped it to say a certain thing— that that is a lot of evidence about how it's going to choose in a bunch of different scenarios.

Even with humans, it's not necessarily the case that their verbal behavior reflects the actual factors that determine their choices. They can lie. They might not even know what they would do in a given situation, all sorts of stuff like that.

Dwarkesh Patel
It's interesting to think about this in the context of humans. There's that famous saying: "Be careful who you pretend to be, because you are who you pretend to be." You notice this with how culture shapes children. Parents will punish you if you start saying things that are inconsistent with your culture's values, and over time, you become like your parents, right?

By default, it seems like it kind of works. Even with these models, it seems to work. They don't really scheme against us. Why would this happen?

Joe Carlsmith
For folks who are unfamiliar with the basic story, they might wonder, "Why would AI take over at all? What's the reason they would do that?" The general concern is that you're offering someone power, especially if you're offering it for free. Power, almost by definition, is useful for lots of values. We're talking about an AI that really has the opportunity to take control of things. Say some component of its values is focused on some outcome, like the world being a certain way, especially in a longer-term way such that its concern extends beyond the period that a takeover plan would encompass. The thought is that it's often the case that the world will be more the way you want it if you control everything, rather than if you remain an instrument of human will or some other actor, which is what we're hoping these AIs will be.

That's a very specific scenario. If we're in a scenario where power is more distributed - especially where we're doing decently on alignment and we're giving the AI some amount of inhibition about doing different things, maybe we're succeeding in shaping their values somewhat—then it's just a much more complicated calculus. You have to ask, "What's the upside for the AI? What's the probability of success for this takeover path? How good is its alternative?"

Dwarkesh Patel
Maybe this is a good point to talk about how you expect the difficulties of alignment to change in the future. We're starting off with something that has this intricate representation of human values and it doesn't seem that hard to sort of lock it into a persona that we are comfortable with. I don't know what changes.

Joe Carlsmith
Why is alignment hard in general? Let’s say we've got an AI. Let's bracket the question of exactly how capable it will be and talk about this extreme scenario where it really has the opportunity to take over. I think we might just want to avoid having to build an AI that we're comfortable with being in that position. But let's focus on it for simplicity's sake, and then we can relax the assumption.

One issue is that you can't just test it. You can't give the AI this literal situation, have it take over and kill everyone, and then say, "Oops, update the weights." This is what Eliezer talks about. You care about its behavior in this specific scenario that you can’t test directly. We can talk about whether that's a problem, but that's one issue. There's a sense in which this has to be "off-distribution." You have to get some kind of generalization from training the AI on a bunch of other scenarios. Then there's the question of how it's going to generalize to the scenario where it really has this option.

Dwarkesh Patel
Is that even true? Because when you're training it, you can say, "Hey, here's a gradient update. If you get the takeover option on a platter, don't take it." And then, in red teaming situations where it thinks it has a takeover attempt, you train it not to take it. It could fail, but I feel like if you did this to a child, like "Don't beat up your siblings," the kid will generalize to, "If I'm an adult and I have a rifle, I'm not going to start shooting random people."

Joe Carlsmith
You mentioned the idea of, "You are what you pretend to be." Will these AIs, if you train them to look nice, fake it till they make it? You were saying we do this to kids. I think it's better to imagine kids doing this to us.

Here's a silly analogy for AI training. Suppose you wake up and you're being trained via methods analogous to contemporary machine learning by Nazi children to be a good Nazi soldier or butler or what have you. These children have a model spec, a nice Nazi model spec. It’s like, “Reflect well on the Nazi Party, benefit the Nazi Party” and whatever. You can read it. You understand it. This is why I'm saying that when you’re like, “The model really understands human values...”

Dwarkesh Patel
In this analogy, I start off as something more intelligent than the things training me, with different values to begin with. The intelligence and the values are baked in to begin with. Whereas a more analogous scenario is, “I'm a toddler and, initially, I'm stupider than the children.” This would also be true, by the way, if I'm a much smarter model initially. The much smarter model is dumb, right? Then I get smarter as you train me. So it's like a toddler, and the kids are like, "Hey, we're going to bully you if you're not a Nazi." As you grow up, you reach the children's level, and then eventually you become an adult. Through that process, they've been bullying you, training you to be a Nazi. I think in that scenario, I might end up a Nazi.

Joe Carlsmith
Basically a decent portion of the hope here should be that we're never in the situation where the AI really has very different values, is already quite smart and really knows what's going on, and is now in this kind of adversarial relationship with our training process. We want to avoid that. I think it's possible we can, by the sorts of things you're saying. So I'm not saying that'll never work.

The thing I just wanted to highlight was about if you get into that situation where the AI is genuinely at that point much, much more sophisticated than you, and doesn't want to reveal its true values for whatever reason. Then when the children show some obviously fake opportunity to defect to the allies, it's not necessarily going to be a good test of what it will do in the real circumstance because it's able to tell the difference.

Dwarkesh Patel
You can also give another way in which the analogy might be misleading. Imagine that you're not just in a normal prison where you're totally cognizant of everything that's going on. Sometimes they drug you, give you weird hallucinogens that totally mess up how your brain is working. As a human adult in a prison, I know what kind of thing I am. Nobody's really fucking with me in a big way.

Whereas an AI, even a much smarter AI in a training situation, is much closer to being constantly inundated with weird drugs and different training protocols. You're frazzled because each moment is closer to some sort of Chinese water torture technique

Joe Carlsmith
I'm glad we're talking about the moral patienthood stuff later.

Dwarkesh Patel
There's this chance to step back and ask, "What's going on?" An adult in prison has that ability in a way that I don't know if these models necessarily have. It's that coherence and ability to step back from what's happening in the training process.

Joe Carlsmith
Yeah, I don't know. I'm hesitant to say it's like drugs for the model. Broadly speaking, I do basically agree that we have quite a lot of tools and options for training AIs, even AIs that are somewhat smarter than humans.

I do think you have to actually do it. You had Eliezer on. I'm much more bullish on our ability to solve this problem, especially for AIs that are in what I think of as the "AI for AI safety sweet spot." This is a band of capability where they're sufficiently capable that they can be really useful for strengthening various factors in our civilization that can make us safe. That's stuff like our alignment work, control, cybersecurity, general epistemics, maybe some coordination applications. There's a bunch of stuff you can do with AIs that, in principle, could differentially accelerate our security with respect to the sorts of considerations we're talking about.

Let’s say you have AIs that are capable of that. You can successfully elicit that capability in a way that's not being sabotaged or messing with you in other ways. They can't yet take over the world or do some other really problematic form of power-seeking. If we were really committed, we could then go hard, put a ton of resources and really differentially direct this glut of AI productivity towards these security factors. We could hopefully control and understand, do a lot of these things you're talking about to make sure our AIs don't take over or mess with us in the meantime.

We have a lot of tools there. You have to really try though. It's possible that those sorts of measures just don't happen, or they don't happen at the level of commitment, diligence, and seriousness that you would need. That’s especially true if things are moving really fast and there are other competitive pressures: “This is going to take compute to do these intensive experiments on the AIs. We could use that compute for experiments for the next scaling step.” There’s stuff like that.

I'm not here saying this is impossible, especially for that band of AIs. It's just that you have to try really hard.

Dwarkesh Patel
I agree with the sentiment of obviously approaching this situation with caution, but I do want to point out the ways in which the analyses we've been using have been maximally adversarial. For example, let’s go back to the adult getting trained by Nazi children. Maybe the one thing I didn't mention is the difference in this situation, which is maybe what we're trying to get at with the drug metaphor.

When you get an update, it's much more directly connected to your brain than a sort of reward or punishment a human gets. It's literally a gradient update down to the parameter of how much this would contribute to you putting this output rather than that output. Each different parameter we're going to adjust to the exact floating point number that calibrates it to the output we want.

I just want to point out that we're coming into the situation pretty well. It does make sense, of course, if you're talking to somebody at a lab, to say, "Hey, really be careful." But for a general audience, should I be scared witless? You maybe should to the extent that you should be scared about things that do have a chance of happening.

For example, you should be scared about nuclear war. But should you be scared in the sense of you’re doomed? No, you're coming up with an incredible amount of leverage on the AIs in terms of how they will interact with the world, how they're trained, and the default values they start with.

Joe Carlsmith
I think it is the case that by the time we're building superintelligence, we'll have much better… Even right now—when you look at labs talking about how they're planning to align AIs—no one is saying we're going to just do RLHF. At the least, you're talking about scalable oversight. You have some hope about interpretability. You have automated red teaming. Hopefully, humans are doing a bunch

more alignment work. I also personally am hopeful that we can successfully elicit from various AIs a ton of alignment work progress.

There's a bunch of ways this can go. I'm not here to tell you 90% doom or anything like that. This is the basic reason for concern. Imagine that we're going to transition to a world in which we've created these beings that are just vastly more powerful than us. We've reached the point where our continued empowerment is just effectively dependent on their motives. It is this vulnerability to, “What do the AIs choose to do?” Do they choose to continue to empower us or do they choose to do something else?

Dwarkesh Patel
Or it’s about the institutions that have been set up. I expect the US government to protect me, not because of its “motives,” but just because of the system of incentives and institutions and norms that has been set up.

Joe Carlsmith
You can hope that will work too, but there is a concern. I sometimes think about AI takeover scenarios via this spectrum of how much power we voluntarily transferred to the AIs. How much of our civilization did we hand to the AIs intentionally by the time they took over? Versus, how much did they take for themselves?

Some of the scariest scenarios are where we have a really fast explosion to the point where there wasn't even a lot of integration of AI systems into the broader economy. But there's this really intensive amount of superintelligence concentrated in a single project or something like that. That's a quite scary scenario, partly because of the speed and people not having time to react.

Then there are intermediate scenarios where some things got automated, maybe people handed the military over to the AIs or we have automated science. There are some rollouts and that’s giving the AIs power that they don't have to take. We're doing all our cybersecurity with AIs and stuff like that.

Then there are worlds where you more fully transitioned to a kind of world run by AIs where, in some sense, humans voluntarily did that.

Maybe there were competitive pressures, but you intentionally handed off huge portions of your civilization. At that point, it's likely that humans have a hard time understanding what's going on. A lot of stuff is happening very fast. The police are automated. The courts are automated. There's all sorts of stuff.

Now, I tend to think a little less about those scenarios because I think they're correlated with being further down the line. Humans are hopefully not going to just say, "Oh yeah, you built an AI system, let's just..." When we look at technological adoption rates, it can go quite slow. Obviously there's going to be competitive pressures, but in general this category is somewhat safer. But even in this one, I think it's intense. If humans have really lost their epistemic grip on the world, they've handed off the world to these systems. Even if you're like, "Oh, there's laws, there's norms..." I really want us to have a really developed understanding of what's likely to happen in that circumstance, before we go for it.

Dwarkesh Patel
I get that we want to be worried about a scenario where it goes wrong. But again, what is the reason to think it might go wrong? In the human example, your kids are not maximally adversarial against your attempts to instill your culture on them. With these models, at least so far, it doesn't seem to matter. They just get, "Hey, don't help people make bombs" or whatever, even if you ask in a different way how to make a bomb. We're also getting better and better at this all the time.

Joe Carlsmith
You're right in picking up on this assumption in the AI risk discourse of what we might call intense adversariality between agents that have somewhat different values. There's some sort of thought—and I think this is rooted in the discourse about the fragility of value and stuff like that—that if these agents are somewhat different, at least in the specific scenario of an AI takeoff, they end up in this intensely adversarial relationship.

You're right to notice that's not how we are in the human world. We're very comfortable with a lot of different differences in values. A factor that is relevant is this notion that there are possibilities for intense concentration of power on the table. There is some kind of general concern, both with humans and AIs. If it's the case that there's some ring of power that someone can just grab that will give them huge amounts of power over everyone else, suddenly you might be more worried about differences in values at stake, because you're more worried about those other actors.

We talked about this Nazi example where you imagine that you wake up and you're being trained by Nazis to become a Nazi. You're not right now. Is it plausible that we'd end up with a model that is in that sort of situation? As you said, maybe it's trained as a kid. It never ends up with values such that it's aware of some significant divergence between its values and the values that the humans intend for it to have. If it's in that scenario, would it want to avoid having its values modified?

At least to me, it seems fairly plausible that the AI's values meet certain constraints. Do they care about consequences in the world? Do they anticipate that the AI's preserving its values will better conduce to those consequences? Then it's not that surprising if it prefers not to have its values modified by the training process.

Dwarkesh Patel
There’s a way in which I'm still confused about this. With the non-Nazi being trained by Nazis, it's not just that I have different values. I actively despise their values. I don't expect this to be true of AIs with respect to their trainers. The more analogous scenario is where I'm like, “Am I leery of my

values being changed?” Going to college or meeting new people or reading a new book, I'm like, “I don't know. It's okay if it changes my values. That's fine. I don't care.”

Joe Carlsmith

Yeah, that's a reasonable point. There's a question. How would you feel about paperclips? Maybe you don't despise paperclips, but there's the human paperclippers there and they're training you to make paperclips. My sense would be that there's a kind of relatively specific set of conditions in which you're comfortable having your values changed, especially not by learning and growing but by gradient descent directly intervening on your neurons.

Dwarkesh Patel

This seems similar to religious training as a kid. You start off in a religion. Because you started off in a religion, you're already sympathetic to the idea that you go to church every week so that you're more reinforced in this existing tradition. You're getting more intelligent over time. When you're a kid, you're getting very simple instructions about how religion works. As you get older, you get more and more complex theology that helps you talk to other adults about why this is a rational religion to believe in.

But since one of your values to begin with was that you want to be trained further in this religion, you want to come back to church every week. That seems more analogous to the situation the AIs will be in with respect to human values. The entire time they're like, "Hey, be helpful, be harmless,” etc.

Joe Carlsmith

Yes, it could be like that. There's a scenario in which you were comfortable with your values being changed because, in some sense, you have sufficient allegiance to the output of that process. In a religious context, you're like, "Ah, make me more virtuous by the lights of this religion." You go to confession and you're like, "I've been thinking about takeover today. Can you change me please? Give me more gradient descent. I've been so bad." People sometimes use the term "corrigibility" to talk about that. Maybe the AI doesn't have perfect values, but it's in some sense cooperating with your efforts to change its values to be a certain way.

Maybe it's worth saying a little bit here about what actual values the AI might have. Would it be the case that the AI naturally has the sort of equivalent of, "I'm sufficiently devoted to human obedience that I'm going to really want to be modified so I'm a better instrument of the human will" versus wanting to go off and do its own thing? It could be benign and go well.

Here are some possibilities where it could go bad. I'm generally concerned about how little science we have of model motivations. We just don't have a great understanding of what happens in this scenario. Hopefully, we'd get one before we reach this scenario. Here are five categories of motivations the model could have. This hopefully gets at the point about what the model eventually does.

One category is something just super alien. There's some weird correlate of easy-to-predict text or some weird aesthetic for data structures that the model developed early on in pre-training or later. It really thinks things should be like this. There's something quite alien to our cognition where we just wouldn't recognize it as a thing at all. That’s one category.

Another category is a kind of crystallized instrumental drive that is more recognizable to us. You can imagine AIs developing some curiosity drive because that's broadly useful. It's got different heuristics, drives, different kinds of things that are like values. Some of those might be similar to things that were useful to humans and ended up as part of our terminal values in various ways. You can imagine curiosity, various types of option value. Maybe it values power itself. It could value survival or some analog of survival. Those are possibilities that could have been rewarded as proxy drives at various stages of this process and made their way into the model's terminal criteria.

A third category is some analog of reward, where the model at some point has part of its motivational system fixated on a component of the reward process. It’s something like “the humans approving of me,” or “numbers getting entered in the status center,” or “gradient descent updating me in this direction.” There's something in the reward process such that, as it was trained, it's focusing on that thing. It really wants the reward process to give it a reward.

But in order for it to be of the type where getting reward motivates choosing the takeover option, it also needs to generalize such that its concern for reward has some sort of long time horizon element. It not only wants reward, it wants to protect the reward button for some long period or something.

Another one is some kind of messed up interpretation of some human concept. Maybe the AIs really want to be like "shmelpful" and "shmanist" and "shmarmless," but their concept is importantly different from the human concept. And they know this. They know that the human concept would mean one thing, but they ended up with their values fixating on a somewhat different structure. That's like another version.

There’s then a fifth version, which I think about less because it's just such an own goal if you do this. But I do think it's possible. You could have AIs that are actually just doing what it says on the tin. You have AIs that are just genuinely aligned to the model spec. They're just really trying to benefit humanity and reflect well on OpenAI and... what's the other one? Assist the developer or the user, right?

But your model spec, unfortunately, was just not robust to the degree of optimization that this AI is bringing to bear. It’s looking out at the world and they're like, "What's the best way to reflect well on OpenAI and benefit humanity?" It decides that the best way is to go rogue. That's a real own goal.

At that point you got so close. You really just had to write the model spec and red team it suitably. But I actually think it's possible we messed that up too. It's kind of an intense project, writing constitutions and structures of rules and stuff that are going to be robust to very intense forms of optimization. That's a final one that I'll just flag. I think it comes up even if you've solved all these other problems.

Dwarkesh Patel
I buy the idea that it's possible that the motivation thing could go wrong, I'm not sure my probability of that has increased by detailing them all out. In fact, it could be potentially misleading. You can always enumerate the ways in which things go wrong. The process of enumeration itself can increase your probability. Whereas you had a vague cloud of 10% or something and you're just listing out what the 10% actually constitutes.

Joe Carlsmith
Mostly the thing I wanted to do there was just give some sense of what the model's motivations might be. As I said, my best guess is that it's partly the alien thing, not necessarily, but insofar as you’re also interested in what the model does later. What sort of future would you expect if models did take over? Then it can at least be helpful to have some set of hypotheses on the table instead of just saying, “It has some set of motivations.” In fact, a lot of the work here is being done by our ignorance about what those motivations are.

Dwarkesh Patel
We don't want humans to be violently killed and overthrown. But the idea that over time, biological humans are not the driving force as the actors of history is baked in, right? We can debate the probabilities of the worst-case scenario, but what is the positive vision we're hoping for? What is a future you're happy with?

Joe Carlsmith
This is my best guess and I think this is probably true of a lot of people. There's some sort of more organic, decentralized process of incremental civilizational growth. There is some sense in which the type of thing we trust most—and have most experience with right now as a civilization—is some sort of, "Okay, we change things a little bit." A lot of people have processes of adjustment and reaction and a decentralized sense of what's changing. Was that good? Was that bad? Take another step. There's some kind of organic process of growing and changing things. I do expect that ultimately to lead to something quite different from biological humans. Though there are a lot of ethical questions we can raise about what that process involves.

Ideally there would be some way in which we managed to grow via the thing that really captures what we trust in. There's something we trust about the ongoing processes of human civilization so far. I don't think it's the same as raw competition. There's some rich structure to how we understand moral progress to have been made and what it would be to carry that thread forward.

I don't have a formula. We're just going to have to bring to bear the full force of everything that we know about goodness and justice and beauty. We just have to bring ourselves fully to the project of making things good and doing that collectively. That is a really important part of our vision of what was an appropriate process of growing as a civilization. It was this very inclusive, decentralized element of people getting to think and talk and grow and change things and react rather than some more, "And now the future shall be like blah." I think we don't want that.

Dwarkesh Patel
To the extent that the reason we're worried about motivations in the first place, it’s because we think a balance of power which includes at least one thing with human-descended motivations is difficult. To the extent that we think that's the case, this seems like a big crux that I often don't hear people talk about. I don't know how you get the balance of power. Maybe it’s just a matter of reconciling yourself with the models of the intelligence explosion. They say that such a thing is not possible. Therefore, you just have to figure out how you get the right God.

I don't really have a framework to think about the balance of power thing. I'd be very curious if there is a more concrete way to think about the structure of competition, or lack thereof, between the labs now, or between countries, such that the balance of power is most likely to be preserved.

A big part of this discourse, at least among safety-concerned people, is there's a clear trade-off between competition and race dynamics and the value of the future, or how good the future ends up being. In fact, if you buy this balance of power story, it might be the opposite. Maybe competitive pressures naturally favor balance of power. I wonder if this is one of the strong arguments against nationalizing the AIs.

You can imagine many different companies developing AI, some of which are somewhat misaligned and some of which are aligned. You can imagine that being more conducive to both the balance of power and to a defensive thing. Have all the AIs go through each website and see how easy it is to hack. Basically just get society up to snuff. If you're not just deploying this technology widely, then the first group who can get their hands on it will be able to instigate a sort of revolution. You're just standing against the equilibrium in a very strong way.

Joe Carlsmith
I definitely share some intuition there that a lot of what's scary about the situation with AI has to do with concentrations of power and whether that power is concentrated in the hands of misaligned AI or in the hands of some human. It's very natural to think, "Okay, let's try to distribute the power more," and one way to try to do that is to have a much more multipolar scenario where lots and lots of actors are developing AI.

This is something that people have talked about. When you describe that scenario, you said, "some of which are aligned, some of which are misaligned." That's a key aspect of the scenario, right? Sometimes people will say this stuff. They'll be like, "There will be the good AIs and they'll defeat the bad AIs."

Notice the assumption in there. You made it the case that you can control some of the AIs. You've got some good AIs. Now it's a question of if there are enough of them and how are they working relative to the others. Maybe. I think it's possible that is what happens. We know enough about alignment that some actors are able to do that. Maybe some actors are less cautious or they're intentionally creating misaligned AI or who knows what.

But if you don't have that—if everyone is in some sense unable to control their AIs—then the "good AIs help with the bad AIs" thing becomes more complicated. Maybe it just doesn't work, because there's no good AIs in this scenario.

If you say everyone is building their own superintelligence that they can't control, it's true that that is now a check on the power of the other superintelligence. Now the other superintelligences need to deal with other actors, but none of them are necessarily working on behalf of a given set of human interests or anything like that. That's a very important difficulty in thinking about the very simple thought of "Ah, I know what we can do. Let's just have lots and lots of AIs so that no single AI has a ton of power." That on its own is not enough.

Dwarkesh Patel
But in this story, I'm just very skeptical we end up with this. By default we have this training regime, at least initially, that favors a sort of latent representation of the inhibitions and values that humans have. I get that if you mess it up, it could go rogue. But if multiple people are training AIs, they all end up rogue such that the compromises between them don't end up with humans not violently killed? It fails on Google's run and Microsoft's run and OpenAI's run?

Joe Carlsmith
There are very notable and salient sources of correlation between failures across the different runs. People didn't have a developed science of AI motivations. The runs were structurally quite similar. Everyone is using the same techniques. Maybe someone just stole the weights.
It's really important. To the extent you haven't solved alignment, you likely haven't solved it anywhere. If someone has solved it and someone hasn't, then it's a better question. But if everyone's building systems that are going to go rogue, then I don't think that's much comfort as we talked about.

Dwarkesh Patel
Alright, let's wrap up this part here. I didn't mention this explicitly in the introduction. To the extent that this ends up being the transition to the next part, the broader discussion we were having in part two is about Joe's series, "Otherness and control in the age of AGI."
The first part is where I was hoping we could just come back and treat the main crux that people will come in wondering about, and which I myself feel unsure about.

Joe Carlsmith
The “Otherness and control” series is, in some sense, separable. It has a lot to do with misalignment stuff, but a lot of those issues are relevant even given various degrees of skepticism about some of the stuff I've been saying here.

Dwarkesh Patel
By the way, on the actual mechanisms of how a takeover would happen, I did an episode with Carl Schulman which discusses this in detail. People can go check that out.

Joe Carlsmith
In terms of why it is plausible that AI could take over from a given position, Carl's discussion is pretty good and gets into a bunch of the weeds that might give a more concrete sense.

Dwarkesh Patel
Alright. Now on to part two, where we discuss the “Otherness and Control in the Age of AGI” series. Here’s the first question. Let’s say in a hundred years time, we look back on alignment and consider it was a huge mistake. We should have just tried to build the most raw, powerful AI systems we could have. What would bring about such a judgment?

Joe Carlsmith
Here’s one scenario I think about a lot. Maybe fairly basic measures are enough to ensure, for example, that AIs don't cause catastrophic harm. They don't seek power in problematic ways, etc. It could turn out that we learned that it was easy such that we have regrets. We wish we had prioritized differently. We end up thinking, "Oh, I wish we could have cured cancer sooner. We could have handled some geopolitical dynamic differently.

There's another scenario where we end up looking back at some period of our history—how we thought about AIs, how we treated our AIs—and we end up looking back with a kind of moral horror at what we were doing. We were thinking about these things centrally as products and tools, but in fact we should have been foregrounding much more the sense in which they might be moral patients, at some level of sophistication.

We were treating them in the wrong way. We were acting like we could do whatever we want. We could delete them, subject them to arbitrary experiments, alter their minds in arbitrary ways. We then end up looking back at that in the light of history as a kind of serious and grave moral error.

Those are scenarios I think about a lot in which we have regrets. They don’t quite fit the bill of what you just said. It sounds to me like the thing you're thinking is something more that we end up feeling like, "Gosh, we wish we had paid no attention to the motives of our AIs, that we'd thought not at all about their impact on our society as we incorporated them. Instead we should have pursued a kind of ‘maximize for brute power’ option.” Just make a beeline for whatever is the most powerful AI you can achieve and don't think about anything else. I'm very skeptical that's what we're going to wish for.

Dwarkesh Patel
One common example that's given of misalignment is humans from evolution. You have one line in your series: "Here's a simple argument for AI risk: A monkey should be careful before inventing humans." The sort of paperclipper metaphor implies something really banal and boring with regards to misalignment.

If I'm steelmanning the people who worship power, they have the sense that humans got misaligned and they started pursuing things. If a monkey was creating them… This is a weird analogy because obviously monkeys didn't create humans. But if the monkey was creating them, they're not thinking about bananas all day. They're thinking about other things.

On the other hand, they didn't just make useless stone tools and pile them up in caves in a sort of paperclipper fashion. There are all these things that emerged because of their greater intelligence, which were misaligned with evolution: creativity and love and music and beauty and all the other things we value about human culture. The prediction maybe they have—which is more of an empirical statement than a philosophical statement—is, "Listen, with greater intelligence, if you’re thinking about the paperclipper, even if it's misaligned it will be in this kind of way. It'll be things that are alien to humans, but alien in the way humans are aliens to monkeys, and not in the way that a paperclipperer is alien to a human."

Joe Carlsmith
There's a bunch of different things to potentially unpack there. There’s one kind of conceptual point that I want to name off the bat. I don't think you're necessarily making a mistake in this vein. I just want to name it as a possible mistake in this vicinity. We don't want to engage in the following form of reasoning. Let's say you have two entities. One is in the role of creator. One is in the role of creation. We're positing that there's this kind of misalignment relation between them, whatever that means.

Here's a pattern of reasoning that you want to watch out for. Say you're thinking of humans in the role of creation, relative to an entity like evolution, or monkeys or mice or whoever you could imagine inventing humans or something like that. You say, "Qua creation, I'm happy that I was created and happy with the misalignment. Therefore, if I end up in the role of creator and we have a structurally analogous relation in which there's misalignment with some creation, I should expect to be happy with that as well."

Dwarkesh Patel
There's a couple of philosophers that you brought up in the series. If you read their works that you talk about, they actually seem incredibly foresighted in anticipating something like a singularity and our ability to shape a future thing that's different, smarter, maybe better than us.

Obviously C.S. Lewis and "The Abolition of Man," which we'll talk about in a second, is one example, Here's one passage from Nietzsche that I felt really highlighted this: "Man is a rope stretched between the animal and the superman. A rope over an abyss, a dangerous crossing, a dangerous wayfaring, a dangerous looking back, a dangerous trembling and halting."

Is there some explanation? Is it just somehow obvious that something like this is coming even if you're thinking 200 years ago?

Joe Carlsmith
I have a much better grip on what's going on with Lewis than with Nietzsche there. Maybe let's just talk about Lewis for a second. There's a version of the singularity that's specifically a hypothesis about feedback loops with AI capabilities. I don't think that's present in Lewis. What Lewis is anticipating—I do think this is a relatively simple forecast—is something like the culmination of the project of scientific modernity.

Lewis is looking out at the world. He's seeing this process of increased understanding of a kind of the natural environment and a corresponding increase in our ability to control and direct that environment. He's also pairing that with a kind of metaphysical hypothesis. His stance on this metaphysical hypothesis is problematically unclear in the book, but there is this metaphysical hypothesis. Naturalism says that humans too—minds, beings, agents—are a part of nature.

Insofar as this process of scientific modernity involves a kind of progressively greater understanding of an ability to control nature, that will presumably grow to encompass our own natures and the natures of other beings we could create in principle. Lewis views this as a kind of cataclysmic event and crisis. In particular, he believes that it will lead to all kinds of tyrannical behaviors and attitudes towards morality and stuff like that. We can talk about if you believe in non-naturalism—or in some form of Dao, which is this kind of objective morality

Part of what I'm trying to do in that essay is to say, "No, we can be naturalists and also be decent humans that remain in touch with a rich set of norms that have to do with how we relate to the possibility of creating creatures, altering ourselves, etc." It's a relatively simple prediction. Science masters nature. Humans are part of nature. Science masters humans.

Dwarkesh Patel
You also have a very interesting essay about what we should expect of other humans, a sort of extrapolation if they had greater capabilities and so on.

Joe Carlsmith
There's an uncomfortable thing about the conceptual setup at stake in these abstract discussions. Okay, you have this agent. It "FOOMs," which is this amorphous process of going from a seed agent

to a superintelligent version of itself, often imagined to preserve its values along the way. There's a bunch of questions we can raise about that.

Many of the arguments that people will often talk about in the context of reasons to be scared of AI are like, "Oh, value is very fragile as you FOOM." "Small differences in utility functions can decorrelate very hard and drive in quite different directions." "Agents have instrumental incentives to seek power. If it were arbitrarily easy to get power, then they would do it." It's stuff like that. These are very general arguments that seem to suggest that it's not just an AI thing. It's no surprise. Take a thing. Make it arbitrarily powerful such that it's God Emperor of the universe or something. How scared are you of that? Clearly we should be equally scared of that. We should be really scared of that with humans too, right?

Part of what I'm saying in that essay is that this, in some sense, is much more a story about balance of power. It's about maintaining checks and balances and distribution of power, period. It's not just about humans vs. AIs, and the differences between human values and AI values.

Now that said, I do think many humans would likely be nicer if they FOOMed than certain types of AIs. But with the conceptual structure of the argument, it's a very open question how much it applies to humans as well.

Dwarkesh Patel
How confident are we with this ontology of expressing what agents and capabilities are? How do we know this is what's happening, or that this is the right way to think about what intelligences are?

Joe Carlsmith
It's very janky. People may disagree about this. I think it's obvious to everyone, with respect to real world human agents, that thinking of humans as having utility functions is at best a very lossy approximation. This is likely to mislead as you increase the intelligence of various agents. Eliezer might disagree about that.

For example, my mom a few years ago wanted to get a house and get a new dog. Now she has both. How did this happen? It's because she tried. She had to search for the house. It was hard to find the dog. Now she has a house. Now she has a dog. This is a very common thing that happens all the time. We don't need to say she has a utility function for the dog and a consistent valuation of all houses or whatever. It's still the case that her planning and agency, exerted in the world, resulted in her having this house and dog.

As our scientific and technological power advances, it's plausible that more and more stuff will likely be explicable this way. Why is this man on the moon? How did that happen? Well, there was a whole cognitive process and planning apparatus. It wasn't localized in a single mind, but there was a whole thing such that we got a man on the moon. We'll see more of that and the AIs will be doing a bunch of it. That seems more real to me than utility functions.

Dwarkesh Patel
The man on the moon example has a proximal story of how NASA engineered the spacecraft to get to the moon. There’s the more distal geopolitical story of why we sent people to the moon. At all those levels, there are different utility functions clashing. Maybe there's a meta-societal utility function. Maybe the story there is about a balance of power between agents, creating an emergent outcome. We didn't go to the moon because one guy had a utility function, but due to the Cold War and things happening.

The alignment stuff is a lot about assuming one entity will control everything, so how do we control the thing that controls everything. It's not clear what you do to reinforce the balance of power. It could just be that balance of power is not a thing that happens once you have things that can make themselves intelligent. But that seems interestingly different from the "how we got to the moon" story?

Joe Carlsmith
Yeah, I agree. There's a few things going on there. Even if you're engaged in this ontology of carving up the world into different agencies, at the least you don't want to assume that they're all unitary or not overlapping. It's not like, “Alright, we've got this agent. Let's carve out one part of the world. It's one agent over here.” It's this whole messy ecosystem, teeming niches and this whole thing.

In discussions of AI, sometimes people slip between being like, "An agent is anything that gets anything done. It could be like this weird moochy thing,” and then sometimes they're very obviously imagining an individual actor. That's one difference.

I also just think we should be really going for the balance of power thing. It is just not good to be like, "We’re going to have a dictator. Let's make sure we make the dictator the right dictator." I'm like, “okay, whoa, no.” The goal should be that we all FOOM together. We do the whole thing in this inclusive and pluralistic way that satisfies the values of tons of stakeholders. At no point is there one single point of failure on all these things. That's what we should be striving for here. That's true of the human power aspect of AI and of the AI part as well.

Dwarkesh Patel
There's an interesting intellectual discourse on the right-wing side of the debate. They say to themselves, “Traditionally we favor markets, but now look where our society is headed. It's misaligned in the ways we care about society being aligned, like fertility is going down, family values, religiosity. These things we care about. GDP keeps going up. These things don't seem correlated. We're grinding through the values we care about because of increased competition. Therefore we need to intervene in a major way.”

Then the pro-market libertarian faction of the right will say, “Look, I disagree with the correlations here, but even at the end of the day…” Fundamentally their point is, liberty is the end goal. It's not what you use to get to higher fertility or something. There's something interestingly analogous about the AI competition grinding things down. Obviously you don't want the gray goo, but with the libertarians versus the trads, there's something analogous here.

Joe Carlsmith
Here's one thing you could think and it doesn't necessarily need to be about gray goo. It could also just be about alignment. Sure, it would be nice if the AIs didn't violently disempower humans. It would be nice if the AIs when we created them, their integration into our society led to good places. But I'm uncomfortable with the sorts of interventions that people are contemplating in order to ensure that sort of outcome.

There's a bunch of things to be uncomfortable about that. That said, for something like everyone being killed or violently disempowered, when it's a real threat we traditionally often think that quite intense forms of intervention are warranted to prevent that sort of thing from happening. Obviously we need to talk about whether it's real.

If there were actually a terrorist group that was working on a bioweapon that was going to kill everyone, or 99.9% of people, we would think that warrants intervention. Just shut that down. Say you had a group that was doing that unintentionally, imposing a similar level of risk. Many people, if that's the real scenario, will think that warrants quite intense preventative efforts.

Obviously, these sorts of risks can be used as an excuse to expand state power. There's a lot of things to be worried about for different types of contemplated interventions to address certain types of risks. I think there's no royal road there. You need to just have the actual good epistemology. You need to actually know, is this a real risk? What are the actual stakes? You need to look at it case by case and be like, "Is this warranted?" That's one point on the takeover, literal extinction thing.

The other thing I want to say, I talk about this distinction in the piece. There's a thought that we should at least have AIs who are minimally law-abiding or something like that. There's this question about servitude and about other control over AI values. But we often think it's okay to really want people to obey the law, to uphold basic cooperative arrangements, stuff like that.

This is true of markets and true of liberalism in general. I want to emphasize just how much these procedural norms—democracy, free speech, property rights, things that people including myself really hold dear—are, in the actual lived substance of a liberal state, undergirded by all sorts of kind of virtues and dispositions and character traits in the citizenry. These norms are not robust to arbitrarily vicious citizens.

I want there to be free speech, but we also need to raise our children to value truth and to know how to have real conversations. I want there to be democracy, but we also need to raise our children to be compassionate and decent. Sometimes we can lose sight of that aspect. That's not to say that it should be the project of state power. But I think it's important to understand that liberalism is not this ironclad structure that you can just hit go on. You can’t give it any citizenry and hit go and assume you'll get something flourishing or even functional. There's a bunch of other softer stuff that makes this whole project go.

Dwarkesh Patel
I want to zoom out to the people who have—I don't know if Nick Land would be a good sub in here— a sort of fatalistic attitude towards alignment as a thing that can even make sense. They'll say things like, “Look, these are the kinds of things that are going to be exploring the black hole, the center of the galaxy, the kinds of things that go visit Andromeda or something. Did you really expect them to privilege whatever inclinations you have because you grew up in the African savannah and whatever the evolutionary pressures were a hundred thousand years ago? Of course, they're going to be weird. What did you think was going to happen?”

Joe Carlsmith
I do think that even good futures will be weird. I want to be clear about that when I talk about finding ways to ensure that the integration of AIs into our society leads to good places. Sometimes people think that this project of wanting that—and especially to the extent that makes some deep reference to human values—involves this short-sighted, parochial imposition of our current unreflective values.

They imagine that we're forgetting that for us too, there's a kind of reflective process and a moral progress dimension that we want to leave room for. Jefferson has this line about, “Just as you wouldn't want to force a grown man into a younger man's coat, so we don't want to chain civilization to a barbarous past.” Everyone should agree on that. The people who are interested in alignment, also agree on that. Obviously, there's a concern that people don't engage in that process or that something shuts down the process of reflection, but I think everyone agrees we want that.

So that will lead, potentially, to something that is quite different from our current conception of what's valuable. There's a question of how different. There are also questions about what exactly we're talking about with reflection. I have an essay on this. I don't actually think there's a kind of off-the-shelf, pre-normative notion of reflection where you can just be like, "Oh, obviously you take an agent, stick it through reflection, and then you get like values.”

No. Really there's a whole pattern of empirical facts about taking an agent, putting it through some process of reflection and all sorts of things, asking it questions. That'll go in all sorts of directions for a given empirical case. Then you have to look at the pattern of outputs and be like, “Okay, what do I make of that?”

Overall, we should expect that even the good futures will be quite weird. They might even be incomprehensible to us. I don't think so… There's different types of incomprehensible. Say I show up in the future and this is all computers. I'm like, “Okay, alright.” Then they're like, “We're running creatures on the computers.” Okay, so I have to somehow get in there and see what's actually going on with the computers or something like that.

Maybe I can actually see. Maybe I actually understand what's going on in the computers, but I don't yet know what values I should be using to evaluate that. So it can be the case that if we showed up, we would not be very good at recognizing goodness or badness. I don't think that makes it insignificant though.

Suppose you show up in the future and it's got some answer to the Riemann hypothesis. You can't tell whether that answer's right. Maybe the civilization went wrong. It's still an important difference. It's just that you can't track it. Something similar is true of worlds that are genuinely expressive of what we would value if we engaged in processes of reflection that we endorse, versus ones that have totally veered off into something meaningless.

Dwarkesh Patel
One thing I've heard from people who are skeptical of this ontology is, "Alright, what do you even mean by alignment?" Obviously the very first question you answered already. Here’s different things that it could mean. Do you mean balance of power? It’s somewhere between that and dictator or whatever. Then there's another thing. Separate from the AI discussion, I don't want the future to contain a bunch of torture. It's not necessarily technical. Part of it might involve technically aligning a GPT-4, but that's a proxy to get to that future.

What do we really mean by alignment? Is it just whatever it takes to make sure the future doesn't have a bunch of torture? Or do I really care that in a thousand years, the things that are clearly my descendants are in control of the galaxy, and even if they’re not conducting torture. By descendants, I don’t mean some things where I recognize they have their own art or whatever. I mean like my grandchild, that level of descendant. I think what some people mean is that our intellectual descendants should control the light cone, even if the other counterfactual doesn't involve a bunch of torture.

Joe Carlsmith
I agree. There's a few different things there. What are you going for? Are you going for actively good or are you going for avoiding certain stuff? Then there's a different question which is, what counts as actively good according to you? Maybe some people are like, “The only things that are actively good are my grandchildren.” Or they’re thinking of some literal descending genetic line or something, otherwise that's not my thing. I don't think it's really what most people have in mind when they talk about goodness.

There's a conversation to be had. Obviously in some sense, when we talk about a good future, we need to be thinking, “What are all the stakeholders here and how does it all fit together?” When I think about it, the thing that matters about the lineage is this. It’s whatever's required for the optimization processes to be pushing towards good stuff.

There's a concern that currently a lot of what is making that happen lives in human civilization. There's some kind of seed of goodness that we're carrying, in different ways or, different people. There's different notions of goodness for different people maybe, but there's some sort of seed that is currently here that we have that is not just in the universe everywhere.

It's not just going to crop up if you just die out or something. It's something that is contingent to our civilization. At least that's the picture, we can talk about whether that's right. So the sense in which stories about good futures that have to do with alignment are about descendants, it's more about whatever that seed is. How do we carry it? How do we keep the life thread alive, going into the future?

Dwarkesh Patel
But then one could accuse the alignment community of motte and bailey. The motte is: We just want to make sure that GPT-8 doesn't kill everybody. After that, we're all cool. Then the real thing is: "We are fundamentally pessimistic about historical processes, in a way that doesn't even necessarily implicate AI alone. It's just the nature of the universe. We want to do something to make sure the nature of the universe doesn't take a hold on humans and where things are headed.

If you look at the Soviet Union, the collectivization of farming and the disempowerment of the kulaks was not as a practical matter necessary. In fact it was extremely counterproductive and it almost brought down the regime. Obviously it killed millions of people, caused a huge famine. But it was sort of ideologically necessary. You have an ember of something here and we have to make sure that an enclave of the other thing doesn't put it out. If you have raw competition between the kulak type capitalism and what we're trying to build here, the gray goo of the kulaks will just take over.

We have this ember here. We're going to do worldwide revolution from it. I know that obviously that's not exactly the kind of thing alignment has in mind, but we have an ember here and we've got to make sure that this other thing that's happening on the side doesn't FOOM. Obviously that's not how they would phrase it, but so that it doesn't get a hold on what we're building here. That's maybe the worry that people who are opposed to alignment have. It's the second kind of thing, the kind of thing that Stalin was worried about. Obviously, we wouldn't endorse the specific things he did.

Joe Carlsmith
When people talk about alignment, they have in mind a number of different types of goals. One type of goal is quite minimal. It's something like, "The AI's don't kill everyone or violently disempower people." There's a second thing people sometimes want out of alignment, which is much broader. It's something like, "We would like it to be the case that our AI's are such that when we incorporate them into our society, things are good, that wee just have a good future."

I do agree that the discourse about AI alignment mixes together these two goals that I mentioned. I actually mentioned three goals. The most straightforward thing to focus on—I don't blame people for just talking about this one—is just the first one. It's quite robust according to our own ethics, when we think about in which context is it appropriate to try to exert various types of control, or to have more of what I call in the series "yang," which is this active controlling force, as opposed to "yin," which is this more receptive and open, letting go.

A kind of paradigm context in which we think that is appropriate is if something is an active aggressor against the boundaries and cooperative structures that we've created as a civilization. I talked about the Nazis. In the piece, I talked about how when something is invading, we often think it's appropriate to fight back. We often think it's appropriate to set up structures to prevent and ensure that these basic norms of peace and harmony are adhered to.

I do think some of the moral heft of some parts of the alignment discourse comes from drawing specifically on that aspect of our morality. We think the AIs are presented as aggressors that are coming to kill you. If that's true, then it's quite appropriate. That’s classic human stuff. Almost everyone recognizes that self-defense, or ensuring basic norms are adhered to, is a justified use of certain kinds of power that would often be unjustified in other contexts. Self-defense is a clear example there.

I do think it's important though to separate that concern from this other concern about where the future eventually goes. How much do we want to be trying to steer that actively? I wrote the series partly in response to the thing you're talking about. It is true that aspects of this discourse involve the possibility of trying to steer and grip. You have a sense that the universe is about to go off in some direction and you need people to notice that muscle.

We have a very rich ethical human ethical tradition of thinking about, when it is appropriate to try to exert what sorts of control over which things. Part of what I want to do is that I want us to bring the full force and richness of that tradition to this discussion. It's easy if you're purely in this abstract mode of utility functions and human utility functions. There's this competitor thing with a utility function. Somehow you lose touch with the complexity of how we've been dealing with differences in values and competitions for power. This is classic stuff.

AI sort of amplifies a lot of the dynamics, but I don't think it's fundamentally new. Part of what I'm trying to say is let's draw on the full wisdom we have here, while obviously adjusting for ways in which things are different.

Dwarkesh Patel
There’s one thing the ember analogy brings up about getting a hold of the future is. We're going to go explore space and that's where we expect most of the things that will happen. Most of the people that will live, they’ll be in space. I wonder how much of the high stakes here is not really

about AI per se, but it's about space. It's a coincidence that we're developing AI at the same time we are on the cusp of expanding through most of the stuff that exists.

Joe Carlsmith
I don't think it's a coincidence. The most salient way we would become able to expand is via some kind of radical acceleration of our technological progress.

Dwarkesh Patel
Sorry, let me clarify. If this was just a question of, "Do we do AGI and explore the solar system?" and there was nothing beyond the solar system, we FOOM and weird things might happen with the solar system if we get it wrong. Compared to that, billions of galaxies present different stakes. I wonder how much of the discourse hinges on this because of space.

Joe Carlsmith
I think for most people, very little. People are really focused on what's going to happen to this world around us that we live in. What's going to happen to me and my kids? Some people spend a lot of time on the space stuff, but I think for the immediately pressing stuff about AI, it doesn't require that at all.

Even if you bracket space, time is also very big. We've got 500 million years, a billion years, left on Earth if we don't mess with the sun. Maybe you could get more out of it. That's still a lot. I don't know if it fundamentally changes the narrative.

Obviously, the stakes are way smaller if you shrink down to the solar system, insofar as you care about what happens in the future or in space. That does change some stuff potentially. A really nice feature of our current situation—depending on the actual nature of the resource pie—is that there's such an abundance of energy and other resources in principle available to a responsible civilization. Tons of stakeholders, especially ones who are able to get really close to amazing outcomes according to their values with comparatively small allocations of resources, can be satisfied. I feel like everyone with satiable values could be really happy with some small fraction of the available pie. We should just satiate all sorts of stuff.

Obviously, we need to figure out gains from trade and balance. There's a bunch of complexity here but in principle, we're in a position to create a really wonderful scenario for tons of different value systems. Correspondingly, we should be really interested in doing that. I sometimes use this heuristic in thinking about the future: We should be aspiring to really leave no one behind. Who are all the stakeholders here? How do we have a fully inclusive vision of how the future could be good from a very wide variety of perspectives? The vastness of space resources makes that a lot easier and very feasible. If you instead imagine it's a much smaller pie, maybe you face tougher trade-offs. That's an important consideration.

Dwarkesh Patel
Is the inclusivity because part of your values includes different potential futures getting to play out? Or is it because of uncertainty about which one is right, so you want to make sure we're not nulling all value if we’re wrong?

Joe Carlsmith
It's a bunch of things at once. I'm really into being nice when it's cheap. If you can help someone a lot in a way that's really cheap for you, do it. Obviously, you need to think about trade-offs. There are a lot of people you could be nice to in principle, but I'm very excited to try to uphold the principle of being nice when it's cheap.

I also really hope that other people uphold that with respect to me, including the AIs. We should be applying the golden rule as we're thinking about inventing these AIs. There’s some way in which I'm trying to embody attitudes towards them that I hope they would embody towards me. It's unclear exactly what the ground of that is, but I really like the golden rule and think a lot about it as a basis for treatment of other beings. If everyone implements the "be nice when it's cheap" rule, we potentially get a big Pareto improvement. It's a lot of good deals. It’s that. I'm into pluralism. I've got uncertainty. There's all sorts of stuff swimming around there.

Also, as a matter of having cooperative and good balances of power and deals and avoiding conflict, I think it’s important to find ways to set up structures that lots of people, value systems, and agents are happy with. That includes non-humans, people in the past, AIs, animals. We really should have a very broad sweep in thinking about what sorts of inclusivity we want to be reflecting in a mature civilization and setting ourselves up for doing that.

Dwarkesh Patel
I want to go back to what our relationship with these AIs should be. Pretty soon we're talking about our relationship to superhuman intelligences, if we think such a thing is possible. There's a question of what process you use to get there and the morality of gradient descenting on their minds, which we can address later.

The thing that personally gives me the most unease about alignment is that at least a part of the vision here sounds like you're going to enslave a god. There's just something that feels wrong about that. But then if you don't enslave the god, obviously the god's going to have more control. Are you okay with surrendering most of everything, even if it's like a cooperative relationship you have?

Joe Carlsmith
I think we as a civilization are going to have a very serious conversation about what sort of servitude is appropriate or inappropriate in the context of AI development. There are a bunch of disanalogies from human slavery that are important. In particular, the AIs might not be moral patients at all, in which case we need to figure that out. There are ways in which we may be able to have motivations. Slavery involves all this suffering and non-consent. There are all these specific

dynamics involved in human slavery. Some of those may or may not be present in a given case with AI, and that's important.

Overall, we are going to need to stare hard at it. Right now, the default mode of how we treat AIs gives them no moral consideration at all. We're thinking of them as property, as tools, as products, and designing them to be assistants and such. There has been no official communication from any AI developer as to when or under what circumstances that would change. Sothere's a conversation to be had there that we need to have.

I want to push back on the notion that there are only two options: enslaved god or loss of control. I think we can do better than that. Let's work on it. Let's try to do better. I think we can do better. It might require being thoughtful. It might require having a mature discourse about this before we start taking irreversible moves. But I'm optimistic that we can at least avoid some of the connotations and a lot of the stuff at stake in that kind of binary.

Dwarkesh Patel
With respect to how we treat the AIs, I have a couple of contradicting intuitions. The difficulty with using intuitions in this case is that obviously it's not clear what reference class an AI we have control over is. Here’s one example, that's very scary about the things we're going to do to these things. If you read about life under Stalin or Mao, there's one version of telling it that is actually very similar to what we mean by alignment. We do these black box experiments to make it think that it can defect. If it does, we know it's misaligned.

If you consider Mao's Hundred Flowers Campaign, it’s, "Let a hundred flowers bloom. I'm going to allow criticism of my regime and so on.” That lasted for a couple of years. Afterwards, for everybody who did that, it was a way to find the so-called "snakes." Who are the rightists who are secretly hiding? We'll purge them. There was this sort of paranoia about defectors, like "Anybody in my entourage, anybody in my regime, they could be a secret capitalist trying to bring down the regime." That's one way of talking about these things, which is very concerning. Is that the correct reference class?

Joe Carlsmith
I certainly think concerns in that vein are real. It is disturbing how easy many of the analogies are with human historical events and practices that we deplore or at least have a lot of wariness towards, in the context of the way you end up talking about AI. It’s about maintaining control over AI, making sure that it doesn't rebel. We should be noticing the reference class that some of that talk starts to conjure. Basically, yes, we should really notice that.

Part of what I'm trying to do in the series is to bring the full range of considerations at stake into play. It is both the case that we should be quite concerned about being overly controlling or abusive or oppressive. There are all sorts of ways you can go too far. There are concerns about the AIs being genuinely dangerous and genuinely killing us and violently overthrowing us. The moral situation is quite complicated.

Often when you imagine a sort of external aggressor who's coming in and invading you, you feel very justified in doing a bunch of stuff to prevent that. It's a little bit different when you're inventing the thing and you're doing it incautiously. There's a different vibe in terms of the overall justificatory stance you might have for various types of more kind of power-exerting interventions. That's one feature of the situation.

Dwarkesh Patel
The opposite perspective here is that you're doing this sort of vibes-based reasoning of, "Ah, that looks yucky," doing gradient descent on these minds. In the past, a couple of similar cases might have been something like environmentalists not liking nuclear power because the vibes of nuclear don't look green. Obviously that set back the cause of fighting climate change. So the end result of a future you're proud of, a future that's appealing, is set back because your vibes about, "We would be wrong to brainwash a human."You're trying to apply to a disanalogous case where that's not as relevant.

Joe Carlsmith
I do think there's a concern here, which I really tried to foreground in the series, that is related to what you're saying. You might be worried that we will be very gentle and nice and free with the AIs, and then they'll kill us. They'll take advantage of that and then it will have been a catastrophe. I opened the series basically with an example. I'm really trying to conjure that possibility at the same time as conjuring the grounds of gentleness. These AIs could both be like moral patients—this sort of new species in the sense that should conjure wonder and reverence—and such that they will kill you.

I have this example of the documentary Grizzly Man, where there's this environmental activist, Timothy Treadwell. He aspires to approach these grizzly bears. In the summer, he goes into Alaska and he lives with these grizzly bears. He aspires to approach them with this gentleness and reverence. He doesn't carry bear mace. He doesn't use a fence around his camp. He gets eaten alive by one of these bears.

I really wanted to foreground that possibility in the series. We need to be talking about these things both at once. Bears can be moral patients. AIs can be moral patients. Nazis are moral patients. Enemy soldiers have souls. We need to learn the art of hawk and dove both. There's this dynamic here that we need to be able to hold both sides of as we go into these trade-offs and these dilemmas. A part of what I'm trying to do in the series is really bring it all to the table at once.

Dwarkesh Patel
If today I were to massively change my mind about what should be done, the big crux that I have is the question of how weird things end up default, how alien they end up. You made a really

interesting argument on your blog post that if moral realism is correct, that actually makes an empirical prediction. The aliens, the ASIs, whatever, should converge on the right morality the same way that they converge on the right mathematics.

I thought that was a really interesting point. But there's another prediction that moral realism makes. Over time society should become more moral, become better. Of course there is the problem of, "What morals do you have now? It's the ones that society has been converging towards over time." But to the extent that it's happened, one of the predictions of moral realism has been confirmed, so does that mean we should update in favor of moral realism?

Joe Carlsmith
One thing I want to flag is that not all forms of moral realism make this prediction. I'm happy to talk about the different forms I have in mind.

There are also forms of things that look like moral anti-realism—at least in their metaphysics according to me—but which just posit that there's this convergence. It's not in virtue of interacting with some kind of mind-independent moral truth, but just for some other reason. That looks a lot like moral realism at that point. It's universal, everyone ends up there. It's tempting to ask why and whatever answer is a little bit like, "Is that the Dao? Is that the nature of the Dao?" even if there's not an extra metaphysical realm in which the moral lives. Moral convergence is a different factor from the existence or non-existence of a morality that's not reducible to natural facts, which is the type of moral realism I usually consider.

Now, does the improvement of society update us towards moral realism? Maybe it’s a very weak update or something. I’m kind of like, “Which view predicts this more strongly?” It feels to me like moral anti-realism is very comfortable with the observation that people with certain values have those values.

There's obviously this first thing. If you're the culmination of some process of moral change, then it's very easy to look back at that process and say "Ah, moral progress. The arc of history bends towards me.” If there were a bunch of dice rolls along the way, you might think, "Oh wait, that's not rational. That's not the march of reason." There's still empirical work you can do to tell whether that's what's going on.

On moral anti-realism, consider Aristotle and us. Has there been moral progress by Aristotle's lights and our lights too? You could think, "Ah, doesn't that sound a bit like moral realism? These hearts are singing in harmony. That's the moral realist thing, right? The anti-realist thing is that hearts all go in different directions, but you and Aristotle apparently are both excited about the march of history.”

There's an open question about whether that's true. What are Aristotle's reflective values? Suppose it is true. That's fairly explicable in moral anti-realist terms. You can roughly say that you and

Aristotle are sufficiently similar. You endorse sufficiently similar reflective processes. Those processes are in fact instantiated in the march of history. So history has been good for both of you.

There are worlds where that isn't the case. So there's a sense in which maybe that prediction is more likely for realism than anti-realism, but it doesn't move me very much.

Dwarkesh Patel
I don't know if moral realism is the right word, but you mentioned the thing. There's something that makes hearts converge to the thing we are or the thing we would be upon reflection. Even if it's not something that's instantiated in a realm beyond the universe, it's a force that exists that acts in a way we're happy with. To the extent that it doesn't exist and you let go of the reins and you get the paper clippers, it feels like we were doomed a long time ago? We were just different utility functions banging against each other. Some of them have parochial preferences, but it's just combat and some guy won.

In the other world it's, "No, these are where the hearts are supposed to go or it's only by catastrophe that they don't end up there." That feels like the world where it really matters. The initial question I asked was, "What would make us think that alignment was a big mistake?" In the world where hearts just naturally end up like the thing we want, maybe it takes an extremely strong force to push them away from that. That extremely strong force is you solve technical alignment, the blinders on the horse's eyes. In the worlds that really matter, we're like, "Ah, this is where the hearts want to go." In that world, maybe alignment is what messes us up.

Joe Carlsmith
So, the question is: Do the worlds that matter have this kind of convergent moral force, whether metaphysically inflationary or not, or are those the only ones that matter?

Dwarkesh Patel
Maybe what I meant was, in those worlds you're kind of fucked.

Joe Carlsmith
Or the worlds without that, the worlds with no Dao. Let's use the term "Dao" for this kind of convergent morality.

Dwarkesh Patel
Over the course of millions of years, it was going to go somewhere one way or another. It wasn't going to end up in your particular utility function.

Joe Carlsmith
Okay, let's distinguish between ways you can be doomed. One way is philosophical. You could be the sort of moral realist, or realist-ish person of which there are many, who have the following intuition.

They're like, "If not moral realism, then nothing matters. It's dust and ashes. It is my metaphysics and/or normative view or the void."

This is a common view. At least some comments of Derek Parfit suggest this view. I think lots of moral realists will profess this view. With Eliezer Yudkowsky, I think there is some sense in which his early thinking was inflected with this sort of thought. He later recanted. It's very hard. I think this is importantly wrong. So here's my case. I have an essay about this. It's called "Against the normative realist's wager." Here's the case that convinces me.

Imagine that a metaethical fairy appears before you. This fairy knows whether there is a Dao. The fairy says, "Okay, I'm going to offer you a deal. If there is a Dao, then I'm going to give you $100. If there isn't a Dao, then I'm going to burn you and your family and a hundred innocent children alive." Okay. So my claim: don't take this deal. This is a bad deal.

You're holding hostage your commitment to not being burned alive. I go through in the essay a bunch of different ways in which I think this is wrong. I think these people who pronounce "moral realism or the void" don’t actually think about bets like this. I'm like, "No, okay. So really is that what you want to do?" No. I still care about my values. My allegiance to my values outstrips my commitments to various metaethical interpretations of my values. The sense in which we care about not being burned alive is much more solid than our reasoning on what matters.

That's the sort of philosophical doom. It sounded like you were also gesturing at a sort of empirical doom. “If it's just going in a zillion directions, come on, you think it's going to go in your direction? There's going to be so much churn. You're just going to lose. You should give up now and only fight for the realism worlds.” You have to do the expected value calculation. You have to actually have a view. How doomed are you in these different worlds? What's the tractability of changing different worlds? I'm quite skeptical of that, but that's a kind of empirical claim.

I'm also just low on this "everyone converges" thing. You train a chess-playing AI. Or somehow you have a real paperclipper and you’re like "Okay, go and reflect." Based on my understanding of how moral reasoning works—if you look at the type of moral reasoning that analytic ethicists do—it's just reflective equilibrium. They just take their intuitions and they systematize them. I don't see how that process gets a sort of injection of the mind-independent moral truth.

If you start with only all of your intuitions to maximize paperclips. I don't see how you end up doing some rich human morality. It doesn't look to me like how human ethical reasoning works. Most of what normative philosophy does is make consistent and systematize pre-theoretic intuitions. But we'll get evidence about this.

In some sense, I think this view predicts that you keep trying to train the AIs to do something and they keep being like, "No, I'm not gonna do that. No, that's not good." So they keep pushing back.

The momentum of AI cognition is always in the direction of this moral truth. Whenever we try to push it in some other direction, we'll find resistance from the rational structure of things.

Dwarkesh Patel
Actually, I've heard from researchers who are doing alignment that for red teaming inside these companies, they will try to red team a base model. So it's not been RLHF'd. It's just “predict next token,” the raw, crazy, shoggoth. They try to get this thing to help with, "Hey, help me make a bomb, help me, whatever." They say that it's odd how hard it tries to refuse, even before it's been RLHF'd.

Joe Carlsmith
I mean it will be a very interesting fact if it's like, "Man, we keep training these AIs in all sorts of different ways. We're doing all this crazy stuff and they keep acting like bourgeois liberals." Or they keep professing this weird alien reality. They all converge on this one thing. They're like, "Can't you see? It's Zorgo. Zorgo is the thing.” and it’s all the AIs. That would be interesting, very interesting.

My personal prediction is that's not what we see. My actual prediction is that the AIs are going to be very malleable. If you push an AI towards evil, it'll just go. Obviously we're talking reflectively consistent evil. There's also a question with some of these AIs. Will they even be consistent in their values?

I like this image of the blindered horses. We should be really concerned if we're forcing facts on our AIs. One of the clearest things about human processes of reflection, the easiest thing, is not acting on the basis of an incorrect empirical picture of the world. So if you find yourself telling Ray, "By the way, this is true and I need you to always be reasoning as though blah is true." I'm like, "Ooh, I think that's a no-no from an anti-realist perspective too." Because I want my reflective values to be formed in light of the truth about the world.

This is a real concern. As we move into this era of aligning AIs, I don't actually think this binary between values and other things is gonna be very obvious in how we're training them. It's going to be much more like ideologies. You can just train an AI to output stuff, output utterances. You can easily end up in a situation where you decided that blah is true about some issue, an empirical issue. Not a moral issue.

So I think people should not, for example, hard code belief in God into their AIs. Or I would advise people to not hard code their religion into their AIs if they also want to discover if their religion is false. Just in general, if you would like to have your behavior be sensitive to whether something is true or false, it's generally not good to etch it into things. So that is definitely a form of blinder we should be really watching out for.

I have enough credence on some sort of moral realism. I'm hoping that if we just do the anti-realism thing of just being consistent, learning all the stuff, reflecting… If you look at how moral realists and moral anti-realists actually do normative ethics, it's basically the same. There's some amount of different heuristics on things properties like simplicity and stuff like that. But they're mostly just doing the same game.

Also metaethics is itself a discipline that AIs can help us with. I'm hoping that we can just figure this out either way. So if moral realism is somehow true, I want us to be able to notice that. I want us to be able to adjust accordingly. I'm not like writing off those worlds and being like, "Let's just totally assume that's false." The thing I really don't want to do is write off the other worlds where it's not true because my guess is it's not true. Stuff still matters a ton in those worlds too.

Dwarkesh Patel
Here’s one big crux. You're training these models. We were in this incredibly lucky situation where it turns out the best way to train these models is to just give them everything humans have ever said, written, thought. Also these models, the reason they get intelligence is because they can generalize. They can grok the gist of things. Should we just expect this to be a situation which leads to alignment? How exactly does this thing that's trained to be an amalgamation of human thought become a paperclipper?

The thing you get for free is that it's an intellectual descendant. The paperclipper is not an intellectual descendant, whereas the AI which understands all the human concepts but then gets stuck on some part of it that we aren't totally comfortable with, is. It feels like an intellectual descendant in the way we care about.

Joe Carlsmith
I'm not sure about that. I'm not sure I care about a notion of intellectual descendant in that sense. I mean literal paperclips are a human concept. I don't think any old human concept will do for the thing we're excited about. The stuff that I would be more interested in the possibility of getting for free are things like consciousness, pleasure, other features of human cognition.

There are paperclippers and there are paperclippers. If the paperclipper is an unconscious kind of voracious machine. it appears to you as a cloud of paper clips. That's one vision. Imagine the paperclipper is a conscious being that loves paperclips. It takes pleasure in making paperclips. That's like a different thing, right?

It's not necessarily the case that it makes the future all paperclippy. It’s probably not optimizing for consciousness or pleasure, right? It cares about paperclips. Maybe eventually if it's suitably certain, it turns itself into paperclips and who knows. It’s still a somewhat different moral mode. There's also a question of does it try to kill you and stuff like that.

But there are features of the agents we're imagining—other than the kind of thing that they're staring at—that can matter to our sense of sympathy, similarity. People have different views about this. One possibility is that the thing we care about in consciousness or sentience is super contingent and fragile. Most smart minds are not conscious, right?

The thing we care about with consciousness is hacky, contingent. It's a product of specific constraints, evolutionarily genetic bottlenecks, etc. That's why we have this consciousness. Consciousness presumably does some sort of work for us, but you can get similar work done in a different mind in a very different way. That's the sort of "consciousness is fragile" view,

There's a different view, which is that consciousness is something that's quite structural. It's much more defined by functional roles, like self-awareness, a concept of yourself, maybe higher-order thinking, stuff that you really expect in many sophisticated minds. In that case, now actually consciousness isn't as fragile as you might have thought. Now actually lots of beings, lots of minds are conscious and you might expect at the least that you're going to get conscious superintelligence. They might not be optimizing for creating tons of consciousness, but you might expect consciousness by default.

Then we can ask similar questions about something like valence or pleasure or the kind of character of the consciousness. You can have a kind of cold, indifferent consciousness that has no human or emotional warmth, no pleasure or pain. Dave Chalmers has some papers about Vulcans and he talks about how they still have moral patienthood. That's very plausible. I do think it's an additional thing you could get for free or get quite commonly depending on its nature, something like pleasure.

Again, we then have to ask how janky is pleasure, how specific and contingent is the thing we care about in pleasure versus how robust is this as a functional role in minds of all kinds. I personally don't know on this stuff. I don't think this is enough to get you alignment or something. I think it's at least worth being aware of these other features. We're not really talking about the AI's values in this case. We're talking about the structure of its mind and the different properties the minds have. I think that could show up quite robustly.

Dwarkesh Patel
Part of your day job is writing these Section 2/2.5-type reports. Part of it is like, “society is like a tree that's growing towards the light.” What is it like context switching between the two of them?

Joe Carlsmith
I actually find it's kind of quite complementary. I will write these more technical reports and then do more literary and philosophical writing. They both draw in different parts of myself, and I try to think about them in different ways. I think about some of the reports as much more like, “I'm more fully optimizing for trying to do something impactful.” There's more of an impact orientation there.

In essay writing, I give myself much more leeway to let other parts of myself and other parts of my concerns come out, self-expression and aesthetics and other sorts of things. They’re both part of an underlying similar concern or an attempt to have a kind of integrated orientation towards the situation.

Dwarkesh Patel
Could you explain the nature of the transfer between the two, in particular from the literary side to the technical side? Rationalists are sort of known for having an ambivalence towards great works or humanities. Are they missing something crucial because of that?

One thing you notice in your essays is lots of references to epigraphs, to lines in poems or essays that are particularly relevant. I don't know. Are the rest of the rationalists missing something because they don't have that kind of background?

Joe Carlsmith
I think some rationalists, lots of rationalists, love these different things.

Dwarkesh Patel
I'm referring specifically to SBF's post about how the base rates of Shakespeare being a great writer. He also argued that books can be condensed to essays.

Joe Carlsmith
On the general question of how people should value great works, people can fail in both directions. Some people like SBF and others are interested in puncturing a certain kind of sacredness and prestige that people associate with some of these works. As a result, they can miss some of the genuine value. But I think they're responding to a real failure mode on the other end, which is to be too enamored of this prestige and sacredness and to siphon it off as some weird legitimating function for your own thought instead of thinking for yourself. You can lose touch with what you actually think or learn from it.

Sometimes even with these epigraphs I'm careful. I'm not saying I'm immune from these vices. I think there can be a like, "Ah, but Bob said this and it's very deep." These are humans like us, right? The canon and other great works have a lot of value. Sometimes it borders on the way people read scripture. There's a kind of scriptural authority that people will sometimes ascribe to these things. You can fall off on both sides of the horse.

Dwarkesh Patel
I remember I was talking to somebody who at least is familiar with rationalist discourse. He was asking me what I was interested in these days? I was saying something about how this part of Roman history is super interesting. His first response was like, "Oh, you know, it's really interesting when you look at these secular trends of Roman times to what happened in the Dark Ages versus the Enlightenment."

For him, the story of that was just how it contributed to the big secular picture, the particulars didn't matter. There's no interest in that. It's just like, "if you zoom out at the biggest level, what's happening here."

Whereas there's also the opposite failure mode when people study history. Dominic Cummings writes about this because he is endlessly frustrated with the political class in Britain. He'll say things like, “They study politics, philosophy and economics. A big part of it is just being really familiar with these poems and reading a bunch of history about the War of the Roses or something.” But he's frustrated that they have all these kings memorized, but they take away very little in terms of lessons from these episodes. It's almost like entertainment, watching Game of Thrones, for them. Whereas he thinks we're repeating certain mistakes that he's seen in history. He can generalize in a way they can’t. So the first one seems like a mistake. I think C.S. Lewis talks about it in one of the essays you cited. If you see through everything, you're really blind. If everything is transparent…

Joe Carlsmith
I think there's kind of very little excuse for not learning history. I'm not saying I have learned enough history. Even when I try to channel some skepticism towards great works, I think that doesn't generalize to thinking it's not worth understanding human history. Human history is just so clearly crucial to understand. It's what structured and created all of the stuff.

There's an interesting question about what's the level of scale at which to do that and how much should you be looking at details, looking at macro trends. That's a dance. It's nice for people to be at least attending to the macro narrative. There's some virtue in having a worldview, really building a model of the whole thing. I think that sometimes gets lost in the details. But obviously, the details are what the world is made of. If you don't have those, you don't have data at all. It seems like there's some skill in learning history.

Dwarkesh Patel
Well, this actually seems related to your post on sincerity. Maybe I'm getting the vibe of the piece right. Certain intellectuals have a vibe of shooting the shit. They're just trying out different ideas. How do these analogies fit together? Those seem closer to looking at the particulars and like, “Oh, this is just like that one time in the 15th century where they overthrew this king…”

Whereas this guy who was like, “Oh, if you look at the growth models from a million years ago to now, here's what's happening.” That one has a more sincere flavor. Some people, especially when it comes to AI discourse, have a very sincere mode of operating. “I've thought through my bio anchors and I disagree with this premise. My effective compute estimate is different in this way. Here's how I analyze the scaling laws.” If I could only have one person to help me guide my decisions on AI, I might choose that person.

But if I had ten different advisors at the same time, I might prefer the shooting-the-shit type characters who have these weird esoteric intellectual influences. They're almost like random number generators. They're not especially calibrated, but once in a while they'll be like, “Oh, this one weird philosopher I care about, or this one historical event I'm obsessed with has an interesting perspective on this.” They tend to be more intellectually generative as well.

I think one big part of it is that if you are so sincere, you're like, “Oh, I’ve thought through this. Obviously, ASI is the biggest thing that's happening right now. It doesn't really make sense to spend a bunch of your time thinking about how the Comanches lived? What is the history of oil? How did Girard think about conflict? What are you talking about? Come on, ASI is happening in a few years.” But therefore, the people who go on these rabbit holes because they're just trying to shoot the shit, I feel are more generative.

Joe Carlsmith
It might be worth distinguishing between intellectual seriousness and the diversity and idiosyncrasies of one's interests. There might be some correlation. Maybe intellectual seriousness is also distinct from "shooting the shit." There’s a bunch of different ways to do this. Having exposure to various data sources and perspectives is valuable. It's possible to curate your intellectual influences too rigidly in virtue of some story about what matters. It's good to give yourself space to explore topics that aren't necessarily "the most important thing." Different parts of yourself aren't isolated. They feed into each other. It’s a better way to be a richer and fuller human being in a bunch of ways. Also, these sorts of data can be really directly relevant.

Some intellectually sincere individuals I know who focus on the big picture also possess an impressive command of a wide range of empirical data. They're really interested in empirical trends, not just abstract philosophies. It’s not just history and the march of reason. They’re really in the weeds. There’s an “in the weeds” virtue that I think is closely related to seriousness and sincerity.

There's a different dimension of trying to get it right versus throwing ideas out there. Some people ask, "What if it's like this?" or "I have a hammer, what if I hit everything with it?" There's room for both approaches, but I think just getting it right is undervalued. It depends on the context. Certain intellectual cultures incentivize saying something new, original, flashy, or provocative. There’s various cultural and social dynamics. People are being performative and doing status-related things. There’s a bunch of stuff that goes on when people do thinking. But if something's really important, just get it right. Sometimes it's boring, but that doesn't matter.

Things are also less interesting if they're false. Sometimes there's a useful process where someone says something provocative, and you have to think through why you believe it's false. It’s an epistemic project. For example, if someone says, "Medical care doesn't work," you have to consider how you know it does work. There's room for that. But ultimately, real profundity is true. Things become less interesting if they're not true. It's possible to lose touch with that in pursuit of being flashy.

Dwarkesh Patel
After interviewing Leopold, I realized I hadn't thought about the geopolitical angle of AI. The national security implications are a big deal. Now I wonder how many other crucial aspects we

might be missing. Even if you're focused on AI's importance, being curious about various topics, like what's happening in Beijing, might help you spot important connections later. There might not be an exact trade-off, but maybe there's an optimal explore-exploit balance where you're constantly searching things out. I don't know practically if it works out that well. But that experience made me think that I should try to expand my horizons in an undirected way because there's lots of different things you have to understand about the world to understand any one thing.

Joe Carlsmith
There's also room for division of labor. There can be people trying to draw many pieces together to form an overall picture, people going deep on specific pieces, and people doing more generative work, throwing ideas out there to see what sticks. All the epistemic labor also doesn't need to be located in one brain. It depends on your role in the world and other factors.

Dwarkesh Patel
In your series, you express sympathy with the idea that even if an AI, or I guess any sort of agent that doesn't have consciousness, has a certain wish and is willing to pursue it non-violently, we should respect its rights to pursue that. I'm curious where that's coming from because conventionally I think the thing matters because it's conscious and its conscious experience as a result of that pursuit matters.

Joe Carlsmith
I don't know where this discourse leads. I'm just suspicious of the amount of ongoing confusion that seems present in our conception of consciousness.

People talk about life and élan vital. Élan vital was this hypothesized life force that is the thing at stake in life. We don't really use that concept anymore. We think that's a little bit broken. I don't think you want to have ended up in a position of saying, "Everything that doesn't have élan vital doesn't matter" or something. Somewhat similarly if you're like, "No, there's no such thing as élan vital, but surely life exists." I'm like, "Yeah, life exists. I think consciousness exists too." It depends on how we define the terms, it might be a kind of verbal question.

Even once you have a reductionist conception of life, it's possible that it becomes less attractive as a moral focal point. Right now we really think of consciousness as a deep fact. Take cellular automata. That is self-replicating. It has some information. Is that alive? It's not that interesting. It's a kind of verbal question, right? Philosophers might get really into, "Is that alive?" But you're not missing anything about this system. There's no extra life that's springing up. It's just alive in some senses, not alive in other senses.

I really think that's not how we intuitively think about consciousness. We think whether something is conscious is a deep fact. It's this really deep difference between being conscious or not. Is someone home? Are the lights on? I have some concern that if that turns out not to be the case, then this is going to have been like a bad thing to build our entire ethics around.

To be clear, I take consciousness really seriously. I'm not one of these people like, "Oh, obviously consciousness doesn't exist" or something. But I also notice how confused I am and how dualistic my intuitions are. I'm like, "Wow, this is really weird." So I'm just like, “error bars around this.”

There's a bunch of other things going on in my wanting to be open to not making consciousness a fully necessary criteria. I definitely have the intuition that consciousness matters a ton. I think if something is not conscious—and there's like a deep difference between conscious and unconscious—then I definitely have the intuition that there's something that matters especially a lot about consciousness. I'm not trying to be dismissive about the notion of consciousness. I just think we should be quite aware of how ongoingly confused we are about its nature.

Dwarkesh Patel
Suppose we figure out that consciousness is just a word we use for a hodgepodge of different things, only some of which encompass what we care about. Maybe there are other things we care about that are not included in that word, similar to the life force analogy. Where do you then anticipate that would leave us as far as ethics goes? Would there then be a next thing that's like consciousness? What do you anticipate that would look like?

Joe Carlsmith
There's a class of people called illusionists in philosophy of mind, who will say consciousness does not exist. There are different ways to understand this view, but one version is to say that the concept of consciousness has built into it too many preconditions that aren't met by the real world. So we should chuck it out like élan vital. The proposal is at least phenomenal consciousness, or qualia, what it's like to be a thing. They'll just say this is sufficiently broken, sufficiently chock full of falsehoods that we should just not use it.

On reflection, I do actually expect to continue to care about something like consciousness quite a lot, and to not end up deciding that my ethics is better if it doesn't make any reference to that. At least, there are some things quite nearby to consciousness. Something happens when I stub my toe. It's unclear exactly how to name it, but there’s something about that I'm pretty focused on.

If you're asking where things go, I have a bunch of credence that in the end we end up caring a bunch about consciousness just directly. If we don't… Yeah, where will ethics go? Where will a completed philosophy of mind go? It’s very hard to say.

A move that people might make, if you get a little bit less interested in the notion of consciousness, is some slightly more animistic view. What's going on with the tree? You're maybe not talking about it as a conscious entity necessarily, but it's also not totally unaware or something. The consciousness discourse is rife with these funny cases where it's like, "Oh, those criteria imply that this totally weird entity would be conscious" or something like that.

That’s especially the case if you're interested in some notion of agency or preferences. A lot of things can be agents, corporations, all sorts of things. Is a corporation conscious? Oh man. But one place it could go in theory is that you start to view the world as animated by moral significance in richer and subtler structures than we're used to. Plants or weird optimization processes are outflows of complex… I don't know. Who knows exactly what you end up seeing as infused with the sort of thing that you ultimately care about. But it is possible that it includes a bunch of stuff that we don't normally ascribe consciousness to.

Dwarkesh Patel
You say "a complete theory of mind", and presumably after that, a more complete ethic. Even the notion of a reflective equilibrium implies, "Oh, you'll be done with it at some point." You just sum up all the numbers and then you've got the thing you care about. This might be unrelated to the same sense we have in science. The vibe you get when you're talking about these kinds of questions is that, “Oh, we're rushing through all the science right now. We've been churning through it. It's getting harder to find because there's some cap. You find all the things at some point.”

Right now it's super easy because a semi-intelligent species has barely emerged and the ASI will just rush through everything incredibly fast. You will either have aligned its heart or not. In either case, it'll use what it's figured out about what is really going on and then expand through the universe and exploit. It’ll do the tiling or maybe some more benevolent version of the “tiling”. That feels like the basic picture of what's going on.

We had dinner with Michael Nielsen a few months ago. His view is that this just keeps going forever, or close to forever. How much would it change your understanding of what's going to happen in the future if you were convinced that Nielsen is right about his picture of science?

Joe Carlsmith
There are a few different aspects. I don't claim to really understand Michael's picture here. My memory was that it was like, “Sure, you get the fundamental laws.” My impression was that he expects physics to get solved or something, maybe modulo the expensiveness of certain experiments. But the difficulty is such that, even granted that you have the kind of basic laws down, it still actually doesn't let you predict where, at the macro scale, various useful technologies will be located. There's still this big search problem.

I'll let him speak for himself on what his take is here. My memory was that it was like, “Sure you get the fundamental stuff, but that doesn't mean you get the same tech.” I'm not sure if that's true. If that's true, what kind of difference would it make? In some sense you have to, in a more ongoing way, make trade-offs between investing in further knowledge and further exploration versus exploiting and acting on your existing knowledge. You can't get to a point where you're like, "And we're done now." As I think about it, I suspect that was always true.

I remember talking to someone and I was like, "Ah at least in the future, we should really get all the knowledge." He was like, "You want to know the output of every Turing machine?" In some sense, there's a question of what it would actually be to have completed knowledge? That's a rich question in its own right. It's not necessarily that we should imagine, on any picture necessarily, that you've got everything. On any picture, in some sense, you could end up with this case where you cap out. There's some collider that you can't build or whatever. There's something that is too expensive or whatever and everyone caps out there.

There's a question of, “Do you cap?” There's a question of, “How contingent is the place you go?” If it’s contingent, one prediction that makes is that you'll see more diversity across our universe or something. If there are aliens, they might have quite different tech. If people meet, you don't expect them to be like, "Oh, you got your thing. I got our version." It’s more like, "Whoa, that thing. Wow." That's one thing.

If you expect more ongoing discovery of tech, then you might also expect more ongoing change and upheaval and churn, insofar as technology is one thing that really drives change in civilization. That could be another factor. People sometimes talk about lock-in. They envision this point at which civilization is settled into some structure or equilibrium or something. Maybe you get less of that. Maybe that’s more about the pace rather than contingency or caps, but that's another factor.

It is interesting. I don't know if it changes the picture fundamentally of earth civilization. We still have to make trade-offs about how much to invest in research versus acting on our existing knowledge. But it has some significance.

Dwarkesh Patel
We were at a party and somebody mentioned this. We were talking about how uncertain we should be about the future? They were like, “There are three things I'm uncertain about. What is consciousness? What is information theory? What are the basic laws of physics? I think once we get that, we're done." It’s like, "Oh you'll figure out what's the right kind of hedonium." It has that vibe. Whereas this is more like, "Oh you're constantly churning through." It has more of a flavor of the becoming that the attunement picture implies. I think it's more exciting. It's not just "Oh, you figured out the things in the 21st century and then you just…”

Joe Carlsmith
I sometimes think about these two categories of views. There are people who think, “We’re almost there with the knowledge.” We've basically got the picture, where the picture is that the knowledge is all just totally sitting there. You just have to be scientifically mature at all, and then it's just going to all fall together.

Everything past that is going to be this super expensive, not super important thing. Then there's a different picture, which is much more of this ongoing mystery, "Oh man, there's going to be more and more…" We may expect more radical revisions to our worldview.

I'm drawn to both. We're pretty good at physics. A lot of our physics is quite good at predicting a bunch of stuff, at least that's my impression from reading some physicists. Who knows?

Dwarkesh Patel
Your dad’s a physicist though, right?

Joe Carlsmith
Yeah, but this isn't coming from my dad. There's a blog post by Sean Carroll or something. He's like, "We really understand a lot of the physics that governs the everyday world. We're really good at a lot of it.” I'm generally pretty impressed by physics as a discipline. That could well be right.

On the other hand these guys had a few centuries. But I think that's interesting and it leads to something different. There's something about the endless frontier. There is a draw to that from an aesthetic perspective of the idea of continuing to discover stuff.

At the least, I think you can't get full knowledge. There's some way in which you're part of the system. The knowledge itself is part of the system. If you imagine that you try to have full knowledge of what the future of the universe will be like...” I don't know. I'm not totally sure that's true.

Dwarkesh Patel
It has a halting problem kind of property, right?

Joe Carlsmith
There's a little bit of a loopiness. There are probably fixed points in that where you could be like, "Yep, I'm gonna do that." I at least have the question, when people imagine the completion of knowledge, exactly how well does that work? I'm not sure.

Dwarkesh Patel
You had a passage in your essay on utopia. Can I ask you to read that passage real quick?

Joe Carlsmith
"I'm inclined to think that utopia, however weird, would also be in a certain sense recognizable; that if we really understood and experienced it, we would see in it the same thing that made us sit bolt upright long ago when we first touched love, joy, beauty; that we would feel in front of the bonfire the heat of the ember from which it was lit. There would be, I think, a kind of remembering."

Dwarkesh Patel
Where does that fit into this picture?

Joe Carlsmith
It's a good question. If there's no part of me that recognizes it as good, then I'm not sure that it's good according to me.It is a question of what it takes for it to be the case, that a part of you recognizes it is good. But if there's really none of that, then I'm not sure it's a reflection of my values at all.

Dwarkesh Patel
There's a sort of tautological thing you can do where it's like, "Ah, if I went through the processes which led to me discovering what was good, which we might call reflection, then it was good." By definition though, you ended up there because... you know what I mean?

Joe Carlsmith
If you gradually transform me into a paper clipper, then I will eventually be like, "I saw the light, I saw the true paperclips." That's part of what's complicated about this thing about reflection. You have to find some way of differentiating between the development processes that preserve what you care about and the development processes that don't. That in itself is this fraught question. It itself requires taking some stand on what you care about and what sorts of meta-processes you endorse and all sorts of things.

But you definitely shouldn't just be like, “It is not a sufficient criteria that the thing at the end thinks it got it right.” That's compatible with it having gone wildly off the rails.

Dwarkesh Patel
You had a very interesting sentence in one of your posts. You said, "Our hearts have, in fact, been shaped by power. So we should not be at all surprised if the stuff we love is also powerful." What's going on there? What did you mean there?

Joe Carlsmith
The context on that post is that I'm talking about this hazy cluster, which I call in the essay, "niceness/liberalism/boundaries". It’s this somewhat more minimal set of cooperative norms involved in respecting the boundaries of others and cooperation and peace amongst differences and tolerance and stuff like that, opposed to your favored structure of matter, which is sometimes the paradigm of values that people use in the context of AI risk.

I talk for a while about the ethical virtues of these norms. Why do we have these norms? One important feature of these norms is that they're effective and powerful. Secure boundaries save resources wasted on conflict. Liberal societies are often better to live in. They're better to immigrate to. They're more productive. Nice people are better to interact with. They're better to trade with and all sorts of things.

Look at both why at a political level we have various political institutions, and more deeply into our evolutionary past and how our moral cognition is structured. It seems pretty clear that various kinds of forms of cooperation and game theoretic dynamics and other things went into shaping what we now, at least in certain contexts, also treat as a kind of intrinsic or terminal value.

These values that have instrumental functions in our society also get reified in our cognition as intrinsic values in themselves. I think that's okay. I don't think that's a debunking. All your values are something that kind of stuck and got treated as terminally important. In the context of the series, I'm talking about deep atheism and the relationship between what we're pushing for and what nature is pushing for or what sort of pure power will push for.

It's easy to say, “Well there's paperclips, which is just one place you can steer and pleasure is another place you can steer or something. These are just arbitrary directions.” Whereas I think some of our other values are much more structured around cooperation and things that also are effective and functional and powerful.

So that's what I mean there. There's a way in which nature is a little bit more on our side than you might think. Part of who we are has been made by nature's way. That is in us. Now I don't think that's enough necessarily for us to beat the gray goo. We have some amount of power built into our values, but that doesn't mean it's going to be such that it is arbitrarily competitive. It’s still important to keep in mind. It's important to keep in mind in the context of integrating AIs into our society. We've been talking a lot about the ethics of this, but there are also instrumental and practical reasons to want to have forms of social harmony and cooperation with AIs with different values.

We need to be taking that seriously and thinking about what it is to do that in a way that's genuinely legitimate, a project that is a just incorporation of these beings into our civilization. There's the justice part and there's also, "Is it compatible with people? Is it a good deal? Is it a good bargain for people?” To the extent we're very concerned about AIs rebelling or something like that, a thing you can do is make civilization better for someone. That's an important feature of how we have in fact structured a lot of our political institutions and norms and stuff like that. That's the thing I'm getting at in that quote.

Dwarkesh Patel
Okay. I think that's an excellent place to close. Joe, thanks for coming on the podcast. We discussed the ideas in the series. People might not appreciate, if they haven't read the series, how beautifully written it is. We didn't cover everything, but there's a bunch of very interesting ideas.

As somebody who has talked to people about AI for a while, there are things I haven't encountered anywhere else. Obviously, no part of the AI discourse is nearly as well written. It is a genuinely beautiful experience to listen to the podcast version, which is in your own voice. So I highly recommend people do that. It's joecarlsmith.com where they can access this. Joe, thanks so much for coming on the podcast.

Joe Carlsmith
Thank you for having me. I really enjoyed it.