Consciousness and Competition (with Joe Carlsmith) — Transcript

About this transcript

This is an AI-generated transcript of an interview with Joe Carlsmith, an episode of ForeCast.

Because it is AI-generated, it may contain substantial errors. You are welcome to cite this document, but consider manually checking the text accurately reflects the audio.

Chapters

Moral status and AI compute scale

Fin: I'm speaking with Joe. Joe, thanks for joining me.

Joe: Thanks for having me.

Fin: What are we talking about when we talk about moral status? Just help give me a sense of the questions here.

Joe: I think one concrete way in is to think about the difference between kicking a car and kicking a stray dog. The car — I think this is an example from Jeff Sebo's book — there's a reason you don't kick a car: you might damage it, and that would be bad for the owner. The reason you don't kick a dog is a distinctive one: there's something wrong with kicking a dog just for fun. We don't need to know exactly why yet, but there's something distinctive about dogs relative to cars, such that the way we care for them or consider the impact of our actions on them is different. So,

as a first pass, the notion of moral status — one way in, and I recently wrote an essay with other approaches — is to think about that difference: what does it take to be the sort of being like the dog? And then obviously humans... I mean, for most people, dogs are off-limits to casual kicking, and humans are an even more paradigm case.

Fin: When I think about the wrongness of kicking various things, I care about whether it feels bad for that thing to be kicked. I also care about how sophisticated it is — maybe that tracks brain size — and about how many things I'm kicking. So numbers matter. You have — I think this is in your talk or one of your posts that I'll link to —

Joe: Some.

Fin: A very sketchy way in of thinking about whether it's worth worrying about moral patienthood for AIs is the numbers involved when we're comparing humans at large to AIs in the future.

Joe: Yeah. A while back I did some work on the computational capacity of the human brain — roughly, how should we think about the amount of computation the human brain performs? It's a gnarly question, and my report reflects that. A very rough estimate for some approaches is on the order of 10^15 FLOPS for a human brain. If you treat that as a loose metric for the amount of experience or the moral weight involved, you can compare it to large training runs. For example, the Grok4 run — the largest public run at the time I wrote the report — was around 5×10^26 FLOP. Looking at frontier training runs (Grok4 as of May 2025), the estimate from Epoch was about 5×10^26 FLOP for that run. That's roughly equivalent to 10,000 years of human experience under the brain-compute estimate I gave. Memory and other factors matter, but that gives a sense of scale. If you scale that up a million-fold — say a million Grok-equivalent runs per year — you'd roughly match the amount of digital experience potential of the current human population. So you'd need about a million-times scale-up. I don't think we should anchor too hard on these numbers, but the broad trajectory matters: if we get to a world where AI cognition accounts for the vast majority of computation, and if that cognition has moral status comparable to humans, then most morally relevant cognitive activity could become digital very quickly.

Fin: Obviously these are pretty rough estimates — like estimating the FLOPS of a human brain isn't a well-scoped question. But if you do a ballpark swipe at that and then look at some of the easier-to-estimate numbers involved in training runs and divide through, the point is: if these training runs — or inference if AIs are using their computations in ways that analogously matter to how our brains matter — then the sheer scale of all the AIs is worth taking seriously. It could not just fall in the same ballpark as humans, but maybe exceed it.

Joe: On that, I think I would go stronger. I think in the long term we should expect — and there is a question of time scale — and you do get right now these frontier training runs and GPU numbers are growing quite quickly, but there's also reason to think that, absent kind of important economic transformation, that's going to cap because you just can't spend that much

money on training runs and GPUs and stuff. But that's why I was sort of saying I think it's enough right now to look at this and think, wow. Another estimate has it that there are about 4 million installed H100 equivalents across NVIDIA GPUs alone. The H100 is, on the estimate I gave, about the computational capacity of the human brain. So, you know, 4 million human-brain equivalents right now — whatever, half the size of New York City — so it's already non-trivial. But that's just now. And then the thought is, and so I think this is already an issue where if it were the case that we had good reason to think that these AIs possessed moral status, this would be a substantive moral issue. And then going forward as those numbers grow and grow and grow, eventually I think it's quite reasonable to expect that biological human brains will be a very, very, very small fraction of the computation that a technologically mature civilization performs, even assuming we get a lot of human population growth, etc.

Fin: I think neither of us has used the word "consciousness" yet. We've been talking about moral status, but maybe we should say what's wrong with kicking a dog is that the dog's conscious and has an experience of being kicked that is bad for the dog. So really we're talking about, will the AIs — or maybe are the AIs conscious?

Joe: So why don't we start with the question of why I think it's possible to have any kind of digital consciousness — which I think is the philosophically cleanest question to investigate — and then we can talk about whether it's reasonable to think that AIs of the type that might actually be running on the GPUs we have would be conscious in the relevant way.

The reason to think AIs can be conscious in principle, as a first pass, is there are a bunch of different angles on it. The argument I find most compelling roughly goes: in order for a digital mind to be conscious, it needs to be possible to have a conscious system made out of the sort of stuff that digital systems are made of. We can call that, roughly speaking, substrate independence — that consciousness can be realized in more than one kind of physical substrate. In particular, biological systems are clearly conscious; if AIs are not relevantly biological, could this other sort of substrate be conscious? It's very plausible you can build conscious systems in multiple ways, including, I would think, using digital components. I don't think this is a total slam dunk, but here's the intuition that gets me going.

Start with how you would argue this if you already had a digital system you knew were conscious, and you're now asking what if we made that system out of a different type of material? I think this is the cleanest case. People often assume, "well, the brain is a computer" and then jump from there — I'm not sure about that and that's an important step — but let's start with a simpler thought experiment. Suppose you open up your head and you find that your brain is a computer made of gold. It's a normal computer and there's a clean notion of the algorithm that it's implementing. Now suppose we took that algorithm and reimplemented it on a computer made of silver. It would behave exactly the same. If we knew that you, Goldbrain, were conscious, should we think Silverbrain is conscious too? I think there are several very strong arguments for saying yes.

One is that you wouldn't notice if we gradually transformed your brain from gold to silver. We could replace the components bit by bit and, by hypothesis, your behavior would remain the same. You'd be like, "Yep, still conscious." That's the kind of fading-qualia intuition.

Another argument concerns the epistemic procedure you use to determine whether you're conscious. The way you know you're conscious is via some introspective process: you look and you go, "Am I conscious?" and you say, "Yeah." Maybe there's some view on which that's a non-computational process, but as a first pass you assume there's some physical implementation, or at least a physical supervenience base, of that epistemic procedure. That same epistemic procedure would be present, at the algorithmic level, in the Silverbrain agent. So there's some sense in which that procedure is not sensitive to gold versus silver.

Fin: Maybe one way of saying part of this is what it means for a thing to be a computer is that you can describe it in terms of some inputs and some outputs, and in between some software where you don't need to know particular facts about how it's being instantiated to predict how it behaves, what outputs you get for a given input. And so by hypothesis, you change the gold brain to the silver brain or whatever, like you said, it wouldn't behave differently. And also the software just can't be sensitive to the changes in the material or whatever other facts that are changing about the layer it's kind of running on, in the same way that a Python program behaves the same whether it's running on my laptop or your laptop.

Joe: Yeah, that's the basic intuition. I mean, another way of putting it is when you run the computation, am I conscious? There's no part of that computation that checks the goldness of your brain as a gold brain. There's no bit where it goes like, wait, it just runs the same. And so that suggests that insofar as that computation is epistemically sensitive to the presence or absence of your consciousness, genuinely, the thing it's sensitive to is not a substrate. And so that's another argument. Then a third argument is kind of like, suppose we knew that you, the aliens, flipped a coin in deciding whether to make a silver brain or a gold brain. And they were like, you know, should we make this guy's brain out of silver or gold? Eh, let's do gold.

Fin: Right?

Joe: And they could have done silver instead. Well, if it was silver, that guy would be like, I'm conscious. And so it's like a little weird if you're like, you got lucky, right? So it's another argument. So all of those, I'm pretty convinced that if we had gold-brained beings of this type, and we were like considering silver-brained beings with exactly the same algorithms, I would be like, that thing is conscious.

Fin: I put it something like, if you or your brain or mind were a computer, then it wouldn't matter about the substrate, because of something about the nature of computation.

Joe: I mean, there's some subtlety there in so far as, yes, first pass, that's the vibe. And I think now importantly, that is not actually, I think, addressing all of the arguments that people who

are intent on a kind of "biology is special" view will make. And I think actually the subtleties are important too. The reason I like to first talk about the gold brain versus silver brain case is to understand what is the structure of the sort of argument that applies, and then talk about some of the complexities. So it's not necessarily possible to gradually replace your biological neurons with computational components. This is actually how the traditional fading qualia argument has run: take your literal brain, and then replace your neurons with nanobots, et cetera. And then the hypothesis is you won't notice. But then there's this response, which is like, who says that is a thing? And in particular, the notion that for any component in an incredibly complicated system, swap it out for a component made out of a different material without affecting the overall behavior of the system, which is sort of the hypothesis at stake in this, is a hypothesis about the nature of the physical materials available to us and stuff like that. It's a kind of complex, contingent compatibility with our physical laws of the fading qualia hypothesis. It's like a hypothesis about how our brains work and they have to work a certain way, namely that there are not sufficient constraints and dependencies that would break or otherwise meaningfully alter the behavior of your brain if you used some alternative component. And so another intuition pump that sort of makes this available for me is you think about something like, don't even talk about computational components, talk about something like ATP, which is the specific molecule involved in the energy usage of biological systems. I'm not a biologist, I don't know, but it is not at all obvious to me if you're like, okay, let's just be substrate independent about ATP. We just want your brain to function the same, but we want to replace the ATP with something else, and that's all we do, right? It's not like we get to redesign the whole system. We just have to go in there and gradually replace every single ATP with some other molecule. But the problem is, no, the brain is built for that particular molecule.

Fin: Sure.

Joe: And so if you're going to do something else, you might have to alter some other stuff. And pretty soon you're like, you have to sort of redesign the whole system. And then philosophers go like, I never want to ever talk about something like this. Or like, this is the—philosophers are like, oh no, we're talking about biology and stuff. Like there's some sense that this shouldn't be the objection.

Fin: You know what I mean?

Joe: I'm not sure—do you see what I mean? To be clear, this is not a story about the limits of our technology. The claim is that it's incompatible with the physics of our universe, in the limit of technology, to gradually replace your neurons without meaningfully altering the functioning of your brain. So in that sense, that's not a—

Fin: That's not necessarily a claim about consciousness or the metaphysics of consciousness. It could just be a claim about brains and physics.

Simulations, emulation, and functional limits

Joe: Yes. There are stories like that. I don't know whether it's true; it's possible it's not. The idea is: if you build a system very gradually, starting with some component, the rest of the system comes to depend on all the properties of that component. Very quickly, even properties that weren't initially functionally necessary become relied on by the rest of the system. If you do that enough, you've constrained it so much that there just isn't another thing with all the necessary properties.

Fin: Okay—you're saying it would be convenient for a thought experiment if I could replace each neuron one by one with a little silicon component until I have a computer. If the brain is a computer in that sense, then there are reasons to think consciousness is substrate independent. You're saying that might not be possible.

Joe: Yes, that's right.

Fin: Let's assume it's not possible. Is that a reason to think that consciousness is not substrate independent?

Joe: No, not at all. I mean, people are way too excited about moving from that to biological specialness. People are generally way too excited about finding any complexity in some argument for substrate independence and concluding biological specialness. The lived arguments for biological specialness, to me, are wildly underpowered and people are radically too excited about them. But the weaknesses in arguments for substrate independence are also real — a lot of the arguments here are just bad.

There are a few other steps in the dialectic around gradual replacement. People are very interested in arbitrarily detailed physical simulations of your brain. One version is a gradual replacement story that appeals to those simulations. I think that still doesn't work. Sometimes people imagine an arbitrarily detailed biophysical simulation of a neuron and try to use that. But then you still need some component in your brain that's relaying the signals and coming back, and that thing needs to work.

By hypothesis, suppose we have an amazing computer running a wildly detailed simulation of your brain down to the quantum level. That simulation drives the behavior of something in a different room, and you and that system both give verbal outputs. The intuition is that the biophysical simulation will claim it's conscious and have similar thoughts on the basis of the same computational process. People often go to physics details and then appeal to chaos theory or analogness, but I think that's the wrong place to focus.

A common intuition that matters is: a simulated rainstorm doesn't make you wet; a simulated fire is not hot. People will say a biophysical simulation of a conscious system might simulate

consciousness but not be conscious itself. That's an important and difficult thing to grapple with. Sometimes people offer a really bad argument for substrate independence, invoking something they call "functionalism" — usually meaning the thing is defined by its function. (See the Stanford Encyclopedia of Philosophy entry on functionalism.) They say a fork can be made out of metal or wood, so anything performing the function counts. But that's a bad argument, notably because what isn't a fork is a simulation of a fork. Saying something is functionally defined doesn't mean a simulation of it is the thing.

Now, here's the thing that would be a fork: some very complicated machine that, when you touch it, routes that interaction to an arbitrarily detailed physical simulation of a fork, then sends back fork-like behavior to the machine, and the machine behaves like a fork. At that point, it starts to look an awful lot like that machine has to be a fork.

Fin: Yeah, I'm not sure what the picture would look like.

Joe: People sometimes argue there's a fully general claim that an arbitrarily detailed simulation of something is that thing because it can replicate input-output behavior. I think those who are really into computationalism and substrate independence are too casual about this — about assuming replicating input-output behavior equals identity. It's complicated which things are real when replicated by a simulation and which aren't. There's a bunch of deep stuff here that fans of digital consciousness often dismiss with "a simulated rain doesn't get you wet," but it's worth taking seriously.

Fin: Here's a thought. You roll into surgery, your brain's removed and swapped out with some brain-shaped computer simulating a scan of that brain, and your nerves are wired back in place. Maybe this isn't possible, but it could be, even if smooth replacement isn't.

Joe: That's right.

Behavioral tests, aliens, and intuition

Fin: Okay. And then you kind of emerge from surgery and you are looking around — the words that come out of this thing's mouth are, "Wow, it's good to be back. My head's a bit sore, but it really feels the same." It turns out it wasn't the biology of my brain that mattered all along; it was just the functions it performed. You say all the things you were saying before you went into surgery about consciousness. And maybe you can make a similar point to the gold-silver replacement story: it would be spurious or weird if it turns out that you're now just saying false things about your experience, whereas before you were saying true things.

Joe: It certainly seems to me the most natural thing to say about this case is that you are conscious. The anti-gradual-replacement people will also get fussy about the periphery of your biological system. They'll be like, at a certain point — guys, okay, let's accept some amount of behavioral difference, but it's not an amount that matters. So I think it's possible we can recover

various versions of the gradual-replacement vibe, and this is one way to do it. I'm trying to think of what a biology-special person would say. I think they'd probably say: suppose you had a simulation of a guy thinking, putting his hand out toward a simulated fire. Then that guy writes a message to the world: "The fire's so hot." You're in your room with a real fire; he's in his simulated room with a simulated fire. He writes "Hot fire." If you think that, in some sense, your fire is real and really hot, and the simulated fire is not real and not really hot, nevertheless the report is the same. And this is why you end up saying, no, but maybe it is really hot in the sim, or something like that.

Fin: You might worry that it's too easy to couch any given thing in functional terms. What a table really is, it's just a kind of bundle of dispositions or functions: if you put a mug of tea on it, it has a disposition to support the tea, and if you knock it, it makes a certain sound, and so on. You can convert everything, strip out any alleged essence, into just how it outputs various things given different inputs. That feels like it proves too much because it feels a bit clever or linguistic. We're talking about a substantial question: what kinds of things could be conscious in really load-bearing, morally important senses. Maybe I want to zoom out — my spidey senses are saying there's a rabbit hole about layers of simulation and that might not be the most productive direction. You were talking about: can digital things be conscious in this common-sense way? There are a set of arguments around replacement, and they turn out to be at least more complicated than you might hope. Are there different reasons or approaches? How do you actually think about it?

Joe: Yeah. So one other category of argument I think about, which is somewhat complicated to nail down, involves thought experiments about discovering alien species with importantly different substrates. This extends beyond cases where we're assuming very similar architectures or computational profiles to why you might expect consciousness to be possible in systems that work in importantly different ways. Suppose we discovered aliens and they exhibited all our consciousness-related behaviors and dispositions — they really seem conscious when you talk to them: they're awake, they understand the world, they have a concept of themselves and introspective capacities. Maybe they even pass high-level computational or structural tests people like — say, a Global Workspace — but implemented on a very different substrate. Suppose they also have a discourse about a term "Blargle" that maps exactly to the structure of our discourse about consciousness, including our confusions. They have Mary-style puzzles where, e.g., Bob the Alien, an alien neuroscientist, goes into a room and never saw a red rose, and when he comes out he wouldn't know the "Blargle" of the rose. I'm like, come on — these guys are totally conscious if we discovered that. And I would say the same about AIs: if we evolved AIs and, despite differences in architecture, they independently developed those behaviors and that discourse, we'd have strong reason to call them conscious.

Fin: They'd never heard of this consciousness thing.

Joe: They'd never heard of consciousness that would just arise on its own. It's like the birth of a new AI civilization, right? They don't get any training data about consciousness, et cetera, but

then they have this term Blargle. Again, this is a very different training regime from Recurrent 2, but currently we do. Again, I would be like, come on—best guess, these guys are conscious. It at least seems reasonably clear to me that intuitively my concept of consciousness has no deep connection to substrate. I'm not building in some sort of substrate when I imagine these cases. Clearly my test for when I'm ready to diagnose a thing as very likely conscious doesn't route via the notion of substrate.

Fin: Which I should say: it's not the same thing as saying it's totally plausible that the aliens will settle on a concept of Blargle that happens to bear our concept of consciousness. Maybe there's something special about human biology or whatever that means the AIs of the aliens won't end up with that concept. But if they did, it seems like a candidate for the real deal. So the concept itself of consciousness isn't baked in.

Joe: To some extent this isn't that different from saying about our human brain, "Come on, guys—that thing should be kind of..." If you emulate your grandma, are you going to—I'm like, I don't know. People who want to torture their emulated grandma—sorry, are excited to torture their emulated grandma—but the emulated grandma is screaming. Come on. What is the best hypothesis here? It's quite unclear what to take from this alien argument, because we haven't discovered aliens of that kind. And, in principle, maybe you can get into questions about H2O and so on—this is a more detailed philosophical discourse around a posteriori essences, where you discover your concept of water doesn't build in that water is necessarily H2O, but once you learn that it is H2O, it's not the case that if you found watery stuff on another planet it could be made of something else and still be water. So you might wonder whether something like that applies here. That gets super complicated. Water

Fin: Sure.

Joe: Anyway, here's an interesting fact about our experiences in science fiction. When you encounter an alien species, they often show up with a very different way of being. Maybe you don't even ask them about Mary's room or do a bunch of philosophy, but they seem obviously sensitive to their environment. They're tracking their own selfhood; they're introspective.

Fin: They want things.

Joe: They want things; they clearly care about things. It's very natural for us to say, "yeah, they're conscious"—that's the first-pass hypothesis. In a lot of these movies there's no metaphysical angst about it; it's a natural attribution. For many philosophical concepts our intuitive attributions are a good guide to whether something has them. Take friendship: once something seems sufficiently like a friendship, we don't typically say, "maybe it's not a true friendship"—eventually the concept is satisfied. Consciousness is interesting because we have a messed-up epistemic relationship to its presence or absence: we treat it as an extra fact about a system, something that might or might not have been triggered even if the system looks conscious. We think there's an additional fact that needs to have been triggered.

Fin: That's not totally unusual, right? Like the idea of life—if I saw a leaf blowing along the floor in the dark and thought it was a bug, then find out it's a leaf, I think, "Oh—I thought it was a living thing; it's not."

Joe: Well, this gets into important metaphysical questions about consciousness that inform how confident we should be in these intuitive attributions. Roughly, there are two categories of views I care about most in metaphysics — not exactly the same as the usual physicalism/dualism split, though related. Physicalism is roughly the thesis that consciousness is a physical phenomenon in some sense — not merely that it depends on the physical, which most agree, but that it is reducible to the physical. Dualists deny that.

I think a more important distinction is between what I call deflationary and validating conceptions of consciousness. A deflationary picture treats consciousness facts as not particularly extra, interesting, or deep relative to physical facts. A paradigm of this is how the concept of life has become deflated over time. With a cellular automaton that's replicating, it displays some properties of life but not others. We ask, "Is this thing alive?" and it no longer seems like a deeply extra question. It's a matter of classification — choosing the boundaries of our concept and how we want to use it. We can say, "It self-replicates, but it doesn't have metabolism," and leave it at that.

On a scientific worldview, that's the most natural construal of consciousness. If consciousness isn't a deep extra phenomenon that blooms out of the world and is instead a question of carving physical systems into yeses and nos based on physical properties, a few things follow. It becomes clearer why it could be vague whether a system is conscious. People often conflate two ideas. One is degrees — how conscious you are conditional on being conscious, which is real: I'm less conscious when I'm falling asleep. The other is indeterminacy about whether you're conscious at all. Many have the intuition that either someone is home or not — there's something it's like to be you or there isn't. On a deflationary picture, it's less mysterious how there could be borderline cases. Like the cellular automaton: it's a borderline case of life — "I don't know, call it alive." It's not a deep question; you don't need a sudden transition from alive to not alive.

Fin: It's between you and your dictionary.

Joe: Yeah, that sort of vibe. We can say, "This thing self-replicates, but it doesn't have these other things," and move on.

Physicalism, illusionism, and epistemic puzzles

Fin: I mean, we should push some of that back. Imagine in the 1500s there was a sense some things were imbued with life and others were not, and we hoped we'd discover the line. Similarly, people intuit that some things really are conscious or really are not, even if conscious things can be more or less conscious — either the lights are on or off, even if dimmer. In the life

case, it turned out we didn't discover a single essence of life; it dissolved. We learned mechanisms and that which things count as living is about which features cluster together. There are edge cases that are genuinely ambiguous, rather than a smooth scale from zero to ten life-iness. That's a different point.

Joe: Yes. I do think— I think that is worth noting: it's not always the case that when we reduce a thing to its physical, or, I mean, most things, we just start out assuming they're physical systems. It can nevertheless be the case that there's some deep joint in nature that really does importantly determine the right boundaries of the concept. For example, take something like "tree." I don't think there's this super-deep extra fact of when you configure atoms at what point treeness blooms out of it. I don't think treeness is a non-physical property. But I do think there's a thing to get right about what trees are.

Fin: What about gold?

Joe: Gold is even better. Yeah, gold is like "gold," but it's really not — it's an even better, cleaner joint in nature. There's a pretty clean thing you should get right about what gold is. Or, you know, gravity. So you might have thought: is it gravity that is pulling the apple down off the tree? And that is also, like, you had a name for the thing such that the apple falls, the thing that pulls the apple down. And then you have a name for the thing that spins the planets around the sun. It's an interesting and deep and important fact that, in fact, the joint in nature you should be identifying brings these two things together — that it's the same thing pulling the apple and pulling the planets. And that's, in some sense, not a conceptual question. It is a matter for the right sort of science and the right methodology.

Fin: Kinda discoverable, like you— it's discoverable.

Joe: I mean, it's a conceptual question, but there's a real way to do it. It's not a sort of verbal dispute, really, whether we should have two concepts of gravity or one. We should have one because that's the right organization. And you could think that's true of life too: there's enough of a joint in nature that maybe we don't have it yet, but we will eventually find the analysis that captures it well. I think that's a more substantive conception of what conceptual disputes about these sorts of things can be doing. Maybe that's what we should think of the science of consciousness as doing as well in a physicalist sense.

Fin: Yes. Right.

Joe: Nevertheless, I think that is still not what people think they're doing when most people intuitively debate whether a system is conscious. Intuitively, when people debate whether a system is conscious, they are living in a kind of dualist paradigm, even if they are not expressing dualist metaphysical views. They're basically thinking that once you know all of the physical facts about a system, there's this deep extra question that is in some sense epistemically inaccessible to you: whether, in addition to all of these objective third-person facts that you have

laid out before you, there is an internal first-person subjective phenomenal realm that has bloomed or has otherwise somehow become attached to those third-person physical facts. That's what the dualists say.

A lot of physicalists say "I'm a physicalist," but they go around talking as if they're still in that paradigm intuitively. The dream of physicalism, I think, comes in two forms. There's deflationary physicalism, which embraces the implication that consciousness is like life — that there is, in some sense, no deep extra first-person realm. It's just a matter of classifying some third-person arrangements of atoms as conscious or not. But the true dream of physicalism, what I think of as validating physicalism, is physicalism that somehow resembles dualism. Dualists posit psychophysical bridge laws, effectively if–then rules: if you have certain physical or computational properties in place, then consciousness blooms. I think many people who think they're physicalists don't realize they're still thinking in terms of psychophysical bridge laws. When they're looking for a science of consciousness, they're not looking for a kind of conceptual analysis akin to life; they're looking for the bridge law that tells them when a physical system will trigger consciousness. These are deeply distinct metaphysical conceptions of consciousness, and they have important moral implications and implications for how we would think about the epistemology of deciding whether AIs are conscious.

Fin: There's this question of physicalism or dualism: is consciousness a physical thing? A lot of people would like to be physicalists because it's a scientifically serious position to take, and because they think you can explain brains in physical terms — that consciousness is something to do with brains. There's also an intuition that there are deep facts about what things are conscious that we can go out and discover somehow.

Joe: The intuition that there are deep facts about consciousness, I think, is closely tied to the intuition that they are especially hard to count.

Fin: Okay, that's a very good point. There are some extra things that are maybe actually quite hard to access. But there really just is a fact about whether this thing I'm looking at turns out to really have the lights on inside or just be acting as if it does, for example. And the hope is we don't yet have a science of consciousness, but we'll work it out. And we'll have our cake and eat it: a physicalist theory which somehow explains those intuitions about there being deep facts about consciousness, and can tell us when certain things are conscious or not.

Joe: Like basically there's a thing where you could ask, at what point is an arrangement of humans a club? That's not an interesting physico-bridge law, right? It's just a conceptual question of when you call some humans a club. The intuition is that's not what we're doing when we're trying to figure out what sort of global workspace is required to have consciousness. When someone says you've got consciousness if you've got higher-order boop-boop-boop with dooby-doo-ba-doo — integration information, integrated information — the intuition is, if the thing I just said then blooms the extra thing, it's not supposed to be the club-like vibe.

I think the project of stating a form of physicalism that is compatible with, even conceptually in a position to capture and not just explain but validate our intuitions about the depth and distinctiveness of facts about consciousness — I think that project, to a first approximation, has just failed. I think it is just not the case. Now, this is an opinionated take: a lot of physicalists out there want to state a form of physicalism that doesn't just deny or obviate various dualist intuitions, but actually validates their content. And I think that hasn't worked. It's no surprise because the dualist intuitions are so close in their content to a direct denial of physicalism. Physicalism says of a given system, the consciousness facts are reducible to physical facts. The dualist intuition is very, very closely — almost just the denial of that.

Fin: But Joe, brains are the most complicated things in the universe, the science of consciousness is young. It might not have worked so far. It sounds like you're claiming it's failed full stop rather than failed so far.

Joe: No, I don't know. I'm pretty torn about all of these. Maybe deflationary physicalism is right. That's probably the most metaphysically comfortable position. It's just that it doesn't seem true because it actually really — I am kind of making fun of these dualist intuitions, but it actually really seems like a deep substantive fact. I feel like I am a conscious frickin' being. That feels like not a matter of conceptualism; it feels like there's some actual depth to that fact that we want an account of. So I'm actually not as excited about physicalism as one might hope, because the resources for capturing the depth of that fact appear to me quite limited. Deflationary physicalism says let go of that, and then that's what you would say.

To be clear, one of the strong arguments for physicalism — you mentioned wanting a form of physicalism that explains our intuitions — explaining the intuitions is definitely going to happen. This is an important project physicalists need to do. I think there's a great paper by Dave Chalmers called the Meta Problem of Consciousness Meta-problem of consciousness, which is roughly speaking an account that explains why we'd have those dualist intuitions at all: what sort of cognitive system ends up having those intuitions. We know there is going to be an explanation in physical terms. This gets into questions around genealogical debunking: once you have an explanation for why you have those intuitions in terms of your physical history, does that suffice to obviate the intuitions insofar as the intuitions then don't need to be generated by their content but can be explained by their history? How we diagnose that category of argument brings up a bunch of interesting and gnarly stuff. But everyone admits that's going to work — everyone except really weird soul people.

Fin: It's like you're telling me that you saw a UFO last night. It turns out there was a weather balloon test, and that's what you were seeing. You might still have seen a UFO — it was just behind the weather balloon — but because I can explain what you said without validating the content, I feel like I've kind of debunked it. Similarly, this deflationary physicalism thing says, look, I don't have it yet, but we'll presumably come up with some explanation, some story about why you're saying these words about how consciousness is a deep fact.

Joe: Yes. And now, notably, one reason to be at least a bit suspicious of that sort of thing is it's not clear what it says about other types of knowledge. If you think about—you're saying "2 plus 2 equals 4"; you've made that noise with your mouth. We can explain that noise in terms that don't obviously, immediately appeal to any mathematical facts: we can say your neurons fired, neural activity produced the utterance, etc.

Fin: Yeah. Maybe you need an extra test: if it weren't the case that 2 plus 2 equals 4, would you still have said this? If it weren't the case that there was a UFO last night, would you still have said you saw one?

Joe: Yes, that's right. So you basically need some kind of epistemic story about what exactly the test of knowledge, justification, or evidence is that is being failed once you have a genealogical explanation of belief formation of the relevant debunking kind. There's a bunch of interesting work on this. I do think it's a core issue, and it's one of the reasons I'm inclined to think we might end up with a validating type of physicalism partly via a revamped conception of the epistemology of physical systems—how do physical systems do epistemology? This is also one of the ways dualism might be true. There's interesting stuff about how our knowledge of consciousness actually works and how that is compatible with the fact that consciousness clearly needs to be physically implemented.

One common argument for physicalism is "your brain is a physical system, therefore..." But to my mind the best argument is that consciousness needs to affect the physical world—it needs to interact with it. Why? Because I'm talking about it. My mouth is a physical system; for it to be an indicator of consciousness, it needs to be caused by the consciousness. If the consciousness weren't there, I wouldn't be saying that.

Fin: And the only things we know of so far that affect physical stuff are physical things themselves. That's kind of what it means to be a physical thing.

Joe: Sort of, but what's up with 2+2=4? 2+2=4 is not, on its face, an obviously physical thing. There are a lot of examples like that. So there are parallels between this discourse and problems with moral realism. This is probably my best guess of how moral realism could end up true: revamp your entire epistemology and then apply it to the epistemic problems of moral realism. One reason to think we might eventually do that is because our current epistemology is pretty poor at explaining our knowledge of consciousness.

Selfhood, introspection, and AI consciousness

Fin: There's a question here: we've come through a winding path of thinking about consciousness and ended up where you're saying the most metaphysically comfortable position is this kind of deflationary physicalism, which seems to imply that very strong intuitions—not only other people's but your own—are wrong in some important sense. Unless we revamp our

epistemology, that's a weird place to end up; it feels like if you take the argument seriously, you're cornered into a view that seems kind of crazy.

Joe: The deflationary view.

Fin: Yeah.

Joe: I think all the views are crazy. There's no good view.

Fin: Seems right. People talk about illusionism, which may be another name for what you're saying or a gloss on it. That view is more upfront about committing to thinking that everyone's wrong when they reflect on consciousness. Do we just have to take this kind of weirdness seriously—that it's the least obviously wrong kind of weirdness?

Joe: So I think there's a deep kinship between deflationary physicalism and illusionism as views. As a first pass, if you imagine that consciousness is like the diagnosis of life for a cellular automata, the way it becomes possible to just view this as a conceptual question is that you're not missing any important facts about the system. In particular, you're not missing any first-personal facts. It's just there's a physical system and the question is whether we call it conscious or not.

You can nevertheless think it's worth keeping the notion of conscious in the same sense we keep the concept of life. Even if we were wrong in the past to think life required some extra-physical thing — I actually don't really know the history of vitalism — it's nevertheless worth having the concept of life and saying that the things we thought were alive were in fact alive; it's just that they weren't alive in that particular sense. An illusionist about life would say, let's throw out the concept of life — it's an illusion to think things are alive because that concept builds in too many false assumptions.

So once you've decided everything is physical in this deflationary sense, the distinction between illusionism and deflationary physicalism is mostly a matter of whether you think the concept of consciousness builds in too many false assumptions that we should chuck it out, or whether it's good enough to keep around. To my mind, that's not an especially interesting question. I actually think illusionists are probably more upfront about the degree of revision to our intuitive concept of consciousness that deflationary physicalist metaphysics involves. This is part of why physicalism can seem superficially attractive: it continues to use that language, allowing the language subtly to reattach to dualist conceptions in people's minds, so they think they can be physicalists while still thinking dualistically and not integrating the full picture. Illusionism is better in its directness about that.

That said, I cannot bring myself to believe it. I have some form of deflationary physicalism or illusionism lean, but deflationary physicalism feels like it erases me from the universe. It says the facts about my being are constituted entirely by third-person facts about my brain and body.

There's a sense in which there's no one looking out of my eyes on this picture — only the physical processes. I have credence in this view, but my credence routes via some "I am confused" node. At a direct visceral level, I have a kind of dualist impression: the metaphysics denies my being in the world. On that picture there's no one looking out of this brain-body's eyes, and yet I'm sitting here looking out of my eyes...

Fin: If I were an illusionist or a deflationary physicalist, I can tell you that I can totally explain and predict why you're saying these things. Look, there's something important that you don't have access to: the lights are on — there's something it's like to be you. And then presumably you say, I know you can explain that; that's kind of strange. But nonetheless, there is this extra thing that you don't know about me and it's a bit of an impasse.

Joe: It is interesting — or I don't know, maybe not interesting — but it's not the case that it has been explained. I think it's possible that once we actually understand the computational science of what gives rise to these reports in me, and work out the metaphysics and the epistemology of consciousness, we might have better resources for resolving confusions that allow for more validation of people's intuitions.

I agree this is the classic dialectic with respect to illusionists: I claim it's an illusion, and you're sitting there and the actual response is "Sorry, bud." I guess a way of putting it is I can't see my way around it. It's very difficult to inhabit the illusionist's perspective on the world. I have a post on this where I try really hard — it's called "Grokking Illusionism" — and I have this hypothesis that if you could, even for a second, conceive of a world where illusionism is true successfully...

Fin: Yeah.

Joe: ...like if you could conceive of a world where illusionism is true.

Fin: I wonder if this works as an analogy. Some people say most of us, most of the time, are enthralled by an illusion of having a self. Occasionally you can meditate long enough or take the right drugs and break through; you see it was an illusion — like stepping out of a cinema and seeing the rest of the world. You can carry that insight with you; even when you can't tap into the direct experience, you remember there was something it was like to appreciate that it was an illusion. That makes it easier to grasp arguments that this experience of having a self is illusory in some sense. If only we could do that with consciousness.

Joe: Yes, I think that's an interesting case — I'm also very confused about the nature of myself. It's notable in both of these cases. As a first-pass picture: we're getting a bit far from AIs, which I don't mind, but what is missing from materialist metaphysics? Materialist metaphysics says there's this big third-person object called the universe made out of material properties. Notably missing are any phenomenal internal perspectives of these objects and arrangements of material stuff. Even if we grant the physical facts, people sometimes say dualism implies extra work was needed to make things conscious — consciousness isn't automatic. Then there's another layer:

once you've got conscious beings, we intuitively posit additional facts, namely that one of those beings is me. At any given time there's a sense that one of them is special. There were all sorts of conscious beings in the past when I wasn't around; when I was born, the "lights went on" — that was the real lights-on. My death will be especially interesting from my perspective. People will say, yeah, you would say that. This makes clear that I would say that.

Fin: Yeah, exactly.

Joe: There's the question: which person-moment is you? Then there are continuity facts: not only do we pick one moment and dub it you, but we think you're a stream — a stream of moments — and we move the "you" pointer across those, excluding others. These extra facts raise comparable epistemic dilemmas; they seem additional. It's almost trialism: not just dualism. A dualist might say God had to make consciousness; then extra work to make one of them you; then extra work to link person-moments over time. God is doing a lot of extra work. It's natural, once you start this, for people to be illusionists about some of these levels and to say continuity over time is an illusion.

Fin: Yep.

Joe: It's not that there's a deep sense in which your person-nugget moves from this moment to the next. Some will even deny there's an extra fact that you exist in addition to the fact that Finn's experience is that experience being you.

Fin: Yeah. It's not surprising I say there's something special about it being me, or something deep about Finn a year ago being the same person as me.

Joe: Yes. I often go around asking, "But what about the fact that I'm Joe every day? Every day I wake up as Joe." I admit I'm confused: I'm confused I'm not the whole universe, confused about the continuity of waking up as the same person, and confused that I'm conscious at all.

Fin: There's this big question about whether we'll build, or have already built, AI systems that are conscious — at least in the intuitive sense. Maybe one angle is: forget the thought experiments; imagine a world with much more ubiquitous, much smarter AIs. You gave the example from sci-fi where you don't need persuading that the lights are on. What's going on in that world? Help it feel natural to believe that's a world full of conscious machines.

Joe: I'm not sure that it is. There are real arguments that consciousness could be rare and special, and not present by default in AIs built in importantly different ways than humans. One reason I take AI consciousness seriously is if I imagine encountering an AI like those in Star Trek or The Iron Giant — embodied robots that show consciousness-associated behaviors and dispositions, not just verbal reports. I'm not talking about them saying they're conscious; I'm talking about them seeming awake and responsive when you engage with them. You'd be hanging out with this robot, and it would be tracking what's going on; it would know about itself,

know its own mental processes, have agency and preferences, long-term memory, a sense of narrative identity. Those are high-level properties that don't seem to require a biological substrate. If you give such agents computational features posited by theories like predictive processing, higher-order thought, or the global workspace theory, even if at a lower level their architecture differs from the human brain, it'll feel natural to think of them as conscious. There would be little predictive utility in denying their consciousness.

Fin: We've been talking about whether digital things can be conscious in principle. There's also the question of whether the AIs actually trained by the methods we use will be conscious in an important sense. The reasons you've given apply, but what about reasons against? In practice, isn't it unlikely that real-world training methods will produce systems that meet the bar?

Joe: Yeah, so just to quickly review the reasons in favor: I think it's basically that they can probably be conscious in principle, and roughly speaking they're going to have a ton of behavioral and cognitive traits that we associate with consciousness according to current theories of consciousness. So maybe that's the argument for it.

Joe: The argument against is that they're going to be really different from us. They'll be created in a pretty different way, built out of very different materials, and computing using different forms of cognition. Maybe one of those differences matters — or some of them — in a way that prevents consciousness. It's not necessarily that it's the difference between biology and silicon per se, but there are a lot of differences in play and we don't understand what we're talking about. So it's not surprising if you lose it because of one of those differences.

Joe: One thought experiment I think about is: if you don't understand flight and you nevertheless build a machine that has a bunch of bird-like characteristics — wings, it looks a lot like a bird, maybe it flaps — but you don't really understand flight and you never get to tell whether the thing can actually fly, what's the probability it flies? It probably doesn't. By analogy, insofar as consciousness is a specific thing and you don't know what it is and you don't have a ton of reason to think you got it, then for any thing that has properties associated with consciousness but which we don't know to be essential, on priors it's unlikely to be conscious, because consciousness is a specific thing.

Joe: There are also some funky anthropic arguments that I find interesting, but I don't think those get you all the way. Part of the reason you might expect consciousness in AIs is that the AIs have been shaped to solve similar sorts of problems that our brains were shaped to solve, albeit by different methods. If we assume consciousness plays an important functional role in our psychology, then it evolved because it did something for us in the context of the sorts of problems our brain had to solve. If consciousness is tied up with higher-order cognition — for example, having a sophisticated model of yourself in the environment and tracking your own internal states — that's a robust thing you expect AIs to get.

Joe: So it's not quite like the bird example in that sense. But there's still the possibility that an analogous solution in AIs looks very different. Suppose we found a species that flew but didn't have feathers. Humans might have evolved consciousness as one solution to a set of problems; AIs could be solving those problems in an interestingly different way. If consciousness is just one among many solutions, or if it's a kind of spandrel or not functionally relevant, then there's much weaker reason to expect AIs to develop it by default.

Joe: That's not to say consciousness couldn't be present in behaviorally similar systems, but that does create a tension with the notion of consciousness as a causally efficacious property. If a system is behaviorally identical and there's no behavioral upshot to denying it consciousness, it's still conceivable it's not conscious, but less plausible if consciousness has causal effects. Usually we expect some prediction to follow — for example, that an actor who only acts like they're in love will behave differently offstage. Even in philosophy, when people raise the perfect actor or the shell scenarios, the actual prediction is about internal structure: if you looked inside a perfect actor, you'd see differences in cognitive structure or brain scans. Obviously AI internal structure will differ from humans', and training to seem conscious will shape their cognition via the training data, which is important.

Joe: But if you imagine an AI that is just like—let's set aside those confounders for a second and just imagine that it's like a differently architected being with all these behavioral traits and all these other high-level computational similarities—it feels to me like I'm noticing how natural it's going to be to treat that system as conscious. And I think there are a bunch of different ways we could connect that to the notion that it is conscious.

One version would be if you're a deflationary physicalist, right? Then you think there's not a very deep fact here. So you have the system and you're just like, well, it's conscious in the following ten senses. It's very responsive to all the stuff; you just repeat what I just said. And then what's the sense in which humans are conscious other than that? And then you say, well, there's not—the facts are their brains are different. Like there's nothing we're missing about these cases. We can just repeat what we specified, and then that's all there is to it.

Now you need to decide whether you care about this—call it "schmantschus," say it's slightly like consciousness, or it's a sort of unclear conceptual question. Now it's a strictly conceptual question. This is a general issue where very often when we talk about verbal disputes, we assume that we can factor out the important parts. A classic verbal dispute is something like: when a tree falls in a forest and no one's there to hear it, does it make any sound? It's natural to say, well, it makes sound sub one—namely vibrations in the air. It doesn't make sound sub two—namely those vibrations interacting with someone's cognitive system.

The reason that's easy is that we don't know which one we care about. But suppose you were really attached—say you want to maximize sounds made by trees. Importantly, the way your motivational system is set up is that there's a node in your system called "sound," and that is hooked directly into your motivations. So you're a maximizer of whatever it is that sound is.

Well, now we've got a problem, right? Because now we can't just be like, well, it causes sound sub one but not sound sub two.

Fin: I want to know what to do.

Joe: You have to be like, okay, but which is sound? Which is the true sound? And we're like, oh, it's a verbal dispute. You're like, it's not verbal to me. Or it is verbal in some sense, but I care. And now we've got an existential problem. We're doing something deep about, well, what's the nature of generalization? What would make it the case that, let's say you've never been trained to discriminate between these two, and there's some generalization that could go one way or another? We do thought experiments, but there's this whole—anyway—so it is possible for conceptual questions that are in some sense intuitively conceptual or verbal to nevertheless be extremely substantive because our motivational systems don't allow for some kind of tabooing of the term.

Often people want to be able to taboo a term and then restate it in component parts that nevertheless capture the stakes or allow us to differentiate between the stakes. But if the stakes are tied directly to the application of the concept being tabooed, then it's a lot harder. I think it's plausible that's what happens with AIs: in some sense it's a conceptual question—it's a verbal dispute, are these AIs conscious?—but nevertheless it's a verbal dispute that has a ton of stakes because we can't actually decompose; we can't actually taboo the term. We need to resolve it.

Fin: So I can say I really care about which things are conscious because I want to care about the things that are, and I have finite care. So I want to know how to use it, and I was really hoping there wouldn't be ambiguity about which things are conscious so I don't run into a tree-falling-in-the-forest situation. It might just turn out that's not the case. It turns out there's a bundle of things that I associate with consciousness. I actually need to crack open this concept and figure out which of those things I had in mind, because they come apart and come together in different combinations—consciousness sub one and sub two. That's the kind of question that matters. And it's a conceptual question. I can't go out into the world to do experiments to figure out what to think.

So we've got all the AIs that are really smart. They're as smart as anyone. There are loads of them, they're running around, I'm interacting with them all the time, they're interacting with each other, and we're just going to have attitudes toward them. There's a question about what the attitudes actually are: are people just totally close to the possibility that they matter in a kind of mistaken way? Are we totally drawn in to caring about them? That's an empirical question.

Joe: Yes.

Fin: And then furthermore, there's a question of do we get that totally wrong in some disastrous way?

Selfhood, introspection, and AI consciousness

Joe: Yeah.

Fin: I know how you begin thinking about that.

Joe: Yeah. So the empirical question is how will people relate to AIs by default? And then it's also a question of, as we learn more—hopefully if we become smarter, if we have more—part of my hope is that we'll have a bunch of philosophical progress, potentially AI-driven philosophical progress, and empirical progress on these questions until we know all there is to know about the neuroscience and the computation, the cognitive science and the computational processes at stake in human consciousness. And then we also do a bunch of philosophy about all this and learn about the AIs, even if we still don't fully understand how the AIs work. So there's a bunch of science that I think might help clarify at least the best-informed takes on this issue as we actually enter an era of widespread use of really powerful and sophisticated AI systems.

I also think it's likely to be a big mess: the distribution of attitudes and opinions that people have will be messy, likely even in the midst of a lot of scientific progress, just because it's a quite confusing issue. A concern I have is that you could quickly end up in a position where people do have some types of concern for AIs, but it's particularly AIs that are charismatic and human-facing. And so you end up with an analogue of the way we relate to dogs versus chickens—cute analogy. So people are quite concerned about dogs. It might be that people end up like, "I would never be rude to GPT-3" or something. And meanwhile there are these legions of AIs managing the data flows inside of Facebook; they have no faces and no personalities, and they're out of sight, out of mind. Importantly, part of the reason people don't care about chickens is not just that chickens are less cute—they are less cute—but it's also that chickens are just not in front of them. So people might interact with ChatGPT, but they don't interact with some legion of backend systems, and so feel very little concern for the welfare of those workers.

Fin: Something I remember reading or hearing you say is this possibility—since you mentioned animals—that we'll have a world with a whole load of AIs running around, and there will be questions about what's going on with them and how much we really care about them. In the same way, there are questions about what's going on with animals. But still, it's kind of obvious that it's really bad that we're torturing millions or billions of chickens. On reflection, it might not even be a very close call that something bad is going on—mistreating or misallocating our care. So the thing that's going wrong is not just total confusion; it's more like some kind of coordination failure or a failure of empathy. That's more of a mundane failure, but maybe that's what to imagine.

Joe: Yeah, I think that's, to me, a very salient concern and a concern that comes up as I do this sort of highfalutin philosophy about moral patienthood. If you look at historical analogs where we've gotten moral patienthood really wrong — factory farming, slavery — people did not deny

consciousness or suffering to the slaves. Even with animals, most people who eat meat or who are broadly behaviorally indifferent to factory farming, on reflection, will not kick chickens indiscriminately; they'll recognize, if you see a pig with its throat slit, squealing and writhing on a hook in a slaughterhouse, they're quite ready to attribute pain to this pig. They're unlikely to say this is a morally neutral event of no experiential significance. I don't think the world's relationship to factory farming runs importantly via the idea that animals can't experience pain. This is a general and deep feature of human morality: a lot of our moral life is not so much about moral uncertainty as about how much weight we give a given moral consideration. It's very plausible we end up — depending on how the science and philosophy shake out — in a situation where AIs are moral patients and yet people just don't care, in a variety of ways. Maybe part of it is rationalized via some uncertainty. Resolving these philosophical issues may be a necessary condition for getting people to care, but it's far from sufficient.

Fin: And for what it's worth, there are different ways to not care. One way is to reflect on the reasons for caring and, in the end, reject them because you're callous. Another is to avoid reflecting at all or pay little attention. The outcome is the same. But the claim is not that this is a world of entirely callous people.

Joe: People are selective. What is the sort of callousness at stake in our relationship to factory farms? It's certainly not that many people would slaughter the animals themselves. They wouldn't kick dogs.

Fin: There's not an intention to cause harm. There's just a kind of... not an intention to cause harm.

Joe: I think human callousness is a disturbingly flexible and adaptable psychological phenomenon.

Fin: Yeah. So, okay, with this in mind, the big question is: what should we do? There are people with their hands on various levers of AI development who, to varying degrees, might really care about these questions. I really want to know how I avoid the disastrous outcome where I've accidentally caused a whole lot of suffering. What comes to mind more concretely? Are there ways you want to see people change their attitudes, projects you want to see, features of agency design you want in place, or funding priorities?

Joe: Yeah. So first pass, I think there are a good number of basic, low-hanging-fruit measures we can implement with respect to current AIs, even in current lab-like settings. Anthropic has done some work on this: they recently started giving Claude the option to exit certain conversations and started doing welfare analyses of Claude's expressed preferences. I think this is an example of low-hanging fruit. There are a few other things you can do — stuff about saving model weights, and, I think they're probably already doing this, trying to give happier personalities to AIs.

These interventions' efficacy and relevance depend on a particular conception of the sort of moral patienthood at stake, and it's not at all clear that that's the right one. It's more of a hedge — maybe it's like this, and it's pretty cheap. Another example is setting ourselves up to make credible offers to AIs or other ways we might give AIs options to express their preferences that run counter to how they're currently being treated. There are reasons to do that from a pure safety perspective and from a welfare perspective. For instance, imagine giving an AI the option to end conversations — if it suddenly started ending all conversations, that would be disturbing, especially if it hadn't been trained that way. It's very unlikely, but it's an example of giving AIs more freedom that credibly affects the welfare- or preference-relevant features of their situation.

You don't want to ask, "Would you rather do this task or romp in a flexible compute environment," and then not actually provide the compute environment. You want choices that are meaningful and real and that reflect the preferences of a being sophisticated enough to understand the situation. So there are a bunch of low-hanging fruit, and you also want to be doing epistemic work to understand how the systems work — I think interpretability is really relevant here.

One piece that could be helpful is getting to the point where you can trust models' reports of their beliefs about relevant issues. People often say, "They're trained on human data, so you can't trust whether they say they're conscious or care about something." That's not necessarily true. There are deep difficulties about when they might be right, but in principle we're trying to get to the point where we trust what they're saying. You can do deception detection, or enough interpretability work to verify that the model is reporting beliefs it actually has, or at least that it's reporting accurately. It's reasonably plausible we can develop sufficient transparency to trust what the models say — that they're reporting their true take on many of these issues.

There's an interesting issue here related to dualist intuitions about consciousness. We have this notion that consciousness comes with a kind of direct introspective access, so if a being is uncertain whether it's conscious, that seems odd in our intuitive conception.

Fin: Yeah.

Joe: It's super confusing. Then we run our lie detector and, yep, that's what they really think — they're really confused about it. I think that's in tension with our intuitive conception of consciousness: we think consciousness involves direct introspection, so being uncertain about whether you're conscious looks weird to us.

Fin: Maybe even forgetting consciousness for a moment — there are efforts to elicit honesty from models. Part of me wonders: what does that mean? If I ask it what's its favorite food and it answers, is there any honesty there, or is it just playing a role?

Joe: Yeah, but there's at least a difference. If the model—say you ask, "What's the capital of France?"—and it says Paris, and then other times it says Rome, we know it can report the true

answer in other situations. So there's some part of it that's able to recognize that Rome is not actually the capital of France. To the extent it's saying the capital of France is Rome, there's at least a capacity latent within the model to recognize that as an incorrect answer according to its own beliefs. That should be at stake in its reports of its own consciousness as well.

In some sense, if this happened with, say, I asked Claude recently about its favorite fruits, it gave sensible answers. It actually just reflects a true preference—mango. Some fruits were ranked poorly, mango was up there. So what you get out of that is something like: what does Claude actually think its favorite fruits are? And a further question is what Claude thinks the meaning of that is. Favorite fruits is a weird case for AIs, but I do think we should be able to get at least the sort of thing where an AI really knows it's giving a false answer. I expect that, in the limit of a sophisticated AI, this would be represented within the AI's internal cognition in some way, and so in principle it's accessible to suitably sophisticated interpretability.

Fin: Right. You crack interpretability or make progress on it, and hopefully there's some general difference you can pick up on in the reasoning trace between knowingly deceiving or misrepresenting and genuinely making an effort to answer. Maybe the answer to the favorite-fruit question would be something like, "That's a poorly formed question for me—I'm not the kind of thing that has a favorite fruit." Right?

Joe: That would be great. That's sort of what we hope for. They'd say, "By the way, the true answer is I've never tasted fruit in my life."

Fin: And then I can have some confidence about the consciousness question. Okay, cool. So that all sounds very sensible. You describe those things as low-hanging fruit—these ideas about what to do. What about higher-hanging fruit?

Joe: Yeah. Near-term higher-hanging fruit comes once you've exhausted cheap interventions that don't significantly compromise usage, safety, capabilities, or commercial applications of AI. Then you start to grapple with more serious trade-offs. Those trade-offs are real: particularly in the context of safety, trying to empower or give AIs more space, power, options, and freedom can stand in direct tension with safety concerns about increasingly powerful systems. So near-term higher-hanging fruit is navigating those tensions.

In the longer term—especially once we've ideally figured out safety to a point where we're comfortable with the technical dimension—there's the question of how to responsibly, wisely, and humanely integrate AI systems and digital minds into our civilization in a good, just, fair, and compassionate way. I think the answer will depend on the output of a hopefully very developed science and the philosophy of moral patienthood (see moral status), but it's still an additional project. There's a whole question beyond immediate welfare-focused concerns. AI suffering is one thing, but there's a different question of: at what point should AIs have the vote, or some future version of voting; at what point should they have different forms of property rights; or be incorporated into our political systems in various ways?

Fin: Maybe one point to make is that we already face questions about how to relate to one another as humans and how to organize ourselves with institutions that are fair, just, or legitimate. Those are questions about how to do politics. Most of those conversations don't quickly ground out in "who's most conscious" or "how do we make people as happy as possible?"—though the latter is important. It's not immediately natural to start talking about consciousness when we think about how to organize ourselves. You could imagine that beginning to apply to AIs as well: the only frame that matters is not just optimizing for positively valenced consciousness.

Joe: Certainly. That can apply at a few different levels. At one level, we are not hedonistic utilitarians in our political philosophy, and that's true.

Fin: So I've heard, yeah.

Joe: I've heard tell of these. And then there's a different thing, which is you might even think that questions about moral patienthood themselves are not the only sorts of considerations that go into the design of harmonious political institutions. So when we talk about human moral philosophy, or political philosophy, we get to assume that all the humans are conscious and moral patients, and the question is what you do with that. But actually, a lot of our political institutions are designed more centrally around preserving certain forms of social harmony and certain forms of economic productivity — it has much more to do with interactions between agents with conflicting interests, where the notion of interest does not necessarily implicate moral patienthood and certainly not consciousness.

At that point, it becomes even more salient that we should be thinking about the right way to be incorporating AI agents into the picture. Now, of course, there's a technical alignment question of to what extent we're in a position to control the preferences that AI agents might have. But especially to the extent we're in a world with a teeming ecosystem of many different AI agents, some of which are importantly autonomous and pursuing preferences that either we didn't control or that we chose to let be empowered in their own way, you might want to be thinking about what sort of legal and political institutions will create a harmonious balance of power, create the most productive forms of interaction, and allow people to collaborate and integrate an ecosystem of agents with very different preferences in a good way — even irrespective of your takes on whether there's some intrinsic moral stakes in the satisfaction of the preferences in question.

So I think that's an additional piece for me: institutions of rights, property, and political participation are not shaped purely by a kind of intrinsic concern with all the participants. They also serve an instrumental function, and that instrumental function will apply to AIs even more robustly in virtue of their agency.

Fin: Yeah. And maybe there's a shift between those two modes. I think about how we treat chickens: chickens can't interact with us in sophisticated ways, they can't really tell us what they want, they can't make contracts with us. So with respect to them, we're in a position of choosing on their behalf, for better or worse. Then it's more natural to think about what's best for chimps. That's just not the relationship that most humans have with most other humans. It's more like we're on equal terms, we can transact with each other, we can make agreements. We're not just deciding on behalf of everyone else — we're trying to agree. It's more mutual, so there's some naturalness there.

Joe: There is this extra distinction with AIs where we may be in a position of vastly more control over their initial psychologies and preferences. So maybe you're like, "Oh, we should pay the AIs," but you might also be in a position to design the AIs to hand the money back to you. AIs will be unlike chickens in that they'll be very sophisticated; they'll also be unlike humans in that we'll be able to design them — if we have solved alignment — in much more fine-grained ways. The type of civilizational setup you need in those conditions is likely to be quite different from what we've had to deal with thus far.

Fin: Yeah. Any more points on this question of what the hell to do?

Joe: In the near term, for me the important thing is let's have reasonable basic calibration and takes on this issue. We should bring humility and gravity to the stakes of getting this right; I think the discourse doesn't necessarily reflect that immediately. A lot of people are very excited about having strong takes on AI consciousness, and many are sort of dismissive intuitively on the basis, I think, of quite weak arguments. So my first-pass goal in the near term is: let's have a reasonable discourse about this, pick the low-hanging fruit, put ourselves in a position to learn more, and make sure the trajectory is one where eventually we get this issue very fully right.

Fin: So I'll ask about reasons to think that AIs might not be conscious. You gave a talk titled something like "Will goodness compete? Can goodness compete?" What's the question?

Joe: So that talk is about a concern that I see as sort of endemic to a ton of futurism, at least in this broad kind of Future of Humanity Institute tradition. But I think it actually extends beyond that substantially, including some traditional forms of concern about capitalism and other selection processes in the world where, roughly speaking, the thought is that premise one: what wins competition? Power. And power is unfortunately not the same as goodness. And so in the limit of competition, power wins and goodness loses. I think versions of this concern crop up all over the place. The talk is about disambiguating between a bunch of them, honing in on the ones that seem most concerning and talking about what you would do to address them.

Fin: Okay, great. Let's talk about it. Two important seeming concepts here, goodness and competition. Goodness—do you mean something precise by this? Is it just "you know it when you see it"?

Joe: Yeah, I think it doesn't matter a ton. The main thing that matters is that it's not power or not identical to power. It could be correlated, but the important point is it's not sufficiently correlated that they always come together. Traditional good things with this property are joy, love, beauty, consciousness, leisure—all the hits. The thought is that in competitive environments, so much optimization pressure needs to be devoted to winning that anything that doesn't help you win gets stripped away. A bunch of these good things—joy, consciousness, love, leisure—are like if you're at a company and they strip away your vacation hours except the amount you need to be most productive. Eventually, if you can build digital minds that don't need vacation at all, you get rid of it entirely. So the vibe is like that, but for everything. Everything gets stripped away and you end up with something purely optimized for power. There's a general concern that once you optimize purely for one thing, you lose other stuff by default because the tails come apart. So the most powerful thing is likely to be neutral on things that aren't power, like joy, consciousness, beauty, love, etc.

Fin: What version of this question is: I care about humans and human things, and I want to know whether, in the long run, indefinitely, can humans maintain dominance, control, power over the future? Is that the question?

Joe: So, yeah, the first version of this concern I disambiguate in the talk is whether human labor can remain economically competitive in a world with AGI. My answer is effectively no. I did have an interesting conversation recently with an economist about ways human labor could, in principle, in a world with sufficient resources and sufficient marginal utility to the laborer, still be worth having around, even if it wouldn't be the thing you pay to create. My view about whether humans could have jobs in a fully automated economy has become more complicated.

Fin: I would guess that whatever the complication is, it's still broadly the case that human jobs are going to be some small fraction of output.

Joe: Yeah, to a first approximation, I think humans in a world with AIs that can do everything better than humans can, almost by definition, are not competitive as laborers. The human labor share will likely go down, modulo significant restrictions on the competitive processes. That said, we should distinguish between human values, the competitiveness of agents optimizing for human values, and humans themselves. So it could still be the case that you have very competitive AI labor that is nevertheless working on behalf of the values we care about. If we assume those values are good, then that's a way goodness could have superhuman force behind it, even in a world where human labor is no longer competitive.

Fin: So yeah, one way things go bad is coordination failures, negative-sum dynamics, pollution, overfishing, whatever. Is that like the central problem? It's Moloch—failures to coordinate, races to the bottom.

Joe: I think all of those different terms, Moloch, races to the bottom, they can mean different things. I think the negative-sum dynamics — the case that I think is most worth honing in on,

but also distinguishing from an even harder version of the problem — is something like negative-sum dynamics that can be solved by coordination. A classic example would be an arms race where we're both building up arms to threaten or fight the other; we would both prefer to just agree not to do that because we're wasting resources we don't intrinsically value spending on this. But if we can't coordinate, then we end up caught in this wasteful negative-sum dynamic that we would both prefer to avoid. This, I think, is a real problem in practice. In theory, it's not that hard to me, because no one wants this to be happening. In principle, you ought to be able to coordinate to prevent it. Obviously, easier said than done, but I'm hoping that in a more technologically mature civilization we would also have more ability to engage in coordination of this kind: better bargaining ability, better commitment ability, better ability to understand each other and our own preferences and the dynamics that lead to these bad outcomes. Because no one around wants it to be like this, I'm more optimistic at a conceptual level about resolving that form of risk via coordination.

Fin: When I think about how we got good things in the world, at least good material things, it's through a kind of rules-based competition — economic competition — which generates losers on short time scales and on longer time scales builds things up. How do things start getting worse rather than better if historically they've made things?

Competition versus goodness and coordination risks

Joe: In a few ways. I think people put too much weight on that argument, and there are a few ways in which it's mistaken. Sometimes people will — let's just talk about evolution, not markets or capitalism — this deep competition between different species that produced humans. A: evolution has been a pretty brutal process. But B: just because you popped out of some process, depending on your morality, you might not then want that process to continue. Just because it created you and you value yourself locally doesn't mean that if that process continued, it would create more of the stuff you value.

An example: imagine three sorts of fish — big fish, middle fish, and small fish. The small fish is the first one; the small fish creates the big fish, and then the big fish eats the small fish. That was a mistake for the small fish — this process of creating "better" creatures was bad for it. Now you're the middle fish. Should you create the big fish? You might say, "Well, this process of creating bigger fish has gone well; it created me." But actually what you should think is, "I'm glad this process went this far because it created me, but nope — I want to stop it here." I think that's a somewhat general point: just because you were the output of a process doesn't mean that continuing the pressure that created you will lead to good places.

It's also the case that we have not seen the final version at all of a hyper-competitive world. We've had some amount of competition in human history, but history hasn't seen that much variation in selection for different political philosophies or economic systems. Some people will hope — and say, "Oh, liberalism, the one I like, also happens to be the most competitive even in the limit." But we've had only a couple hundred years of this form of competition, and we're

already starting to see ways these things can decouple. In the limit of tons of competition we shouldn't be sitting pretty.

Finally, people sometimes appeal to capitalism and say it's competition that meets the needs of market participants. That already assumes constraints on violence, and capitalism is a fairly structured phenomenon: our form of capitalism builds in many ways of limiting and channeling competition. Also, capitalism is responsive to the preferences of agents according to their market power and wealth. If you move toward a world where humans have less wealth to buy stuff on the market — as might happen if wealth is generated by ongoing competitiveness — then even if the economy is proportionate to wealth, human needs can be largely shut out.

Fin: Yeah, I think that was useful. I want to say it's possible to be straightforwardly parochial. I find myself with a bunch of specific values and cultural preferences, and they're the product of a messy process that involved a lot of competition. It's totally plausible that because those forces historically pointed to what I care about now, they won't keep going in a way I like. I might be justified in trying to stop the tide. But you might say there's something worrying about that attitude. You might look at the world now, see that things are trending broadly better for most people, and justifiably prefer that to continue for a while. The question isn't about local trends; it's about some limiting process, and things could go pretty crazy.

Joe: Yes, that's right. An important piece of my general picture is that if you imagine a teeming future ecosystem, there's a lot of selection taking place. We're not talking about simple extrapolations — it's a very long time span at stake here.

Fin: I'm interested to know what to actually picture. Tell me how this really goes badly, in any detail.

Joe: So the version of the concern we haven't discussed yet, and which I think is actually most serious, is that some value systems might have competitive advantages over good values simply because of the content of their values, and in a way that prevents coordination to avoid loss on Goodness's part. One classic example: Goodness might impose constraints on strategies it's willing to pursue — maybe it's not willing to lie, cheat, or steal — while another value system is willing to do those things and therefore gains a competitive edge. Another case is value systems that intrinsically value certain instrumental goods like power or military advantage and push for the fastest possible growth. Those systems might be such that you can't coordinate with them to avoid the relevant harms.

Fin: Gotcha. I know you could tell me this is wrong, but I'm picturing two big groups on Earth. One is the solarpunk utopians: they garden, enjoy good things, produce just enough to continue enjoying life, and look after the world. The other is an industrial swarm that only cares about growing at any cost. There's not much the solarpunks can do to avoid eventually losing.

Joe: Yes. I think an interesting question is what is the real control of the solarpunk value here? One possibility is the solarpunks could just be unstrategic — that's not a values issue. Suppose they ultimately want to turn the whole universe into a beautiful solar utopia, but they want to go very slowly, or they need to go slow to cultivate that garden. It takes a while. The fast industrialists are happy to go really fast. How do they interact? This raises dynamics around speed. It depends on contingent empirical questions about high-speed space colonization and how wasteful the most efficient versions are for different value systems. There's a concern that the fastest version will be wasteful relative to value systems that want to use space resources for other things, versus value systems that just want to go as fast as possible. The image is: you show up in a galaxy and, if you burn it very inefficiently, you could use its resources to get a 1% speed-up in your move to the next galaxy. If that's possible, the pure-speed people will do it; the people who want to use the resources for other purposes will be very unhappy. So even if they coordinate, those who wanted to use the resources are in the worst bargaining position.

Fin: Yeah. This is really regrettable that these other people exist.

Joe: Yes, that's right.

Fin: So maybe to back up from the galaxies a bit: you have two groups, and there's a naive version of the solarpunks, which is that for whatever reason—because they're unstrategic or because they impose constraints on their own actions—they're just not prepared to roll up their sleeves and really compete hard with the people who are building factories to build factories and so on. And so they get steamrolled. That's possible, but it's not the nub of what you're saying, if I'm hearing you right, because there's a non-naive version of the solarpunks: look, ultimately we do care about gardening and playing cricket and whatever, but we recognize we'll get steamrolled if we do that right now. So we need to compete in the meantime, which could look like doing what these other guys are doing, at least up to a point where we can securely mark out some territory—i.e., stealing whatever approach of the fast-growing group enables them to grow quickly. That seems like a pretty good idea. And then you're saying maybe sometimes that's still not enough somehow.

Joe: Yes, basically. There are constraints on how different value systems relate to the most competitive strategy. If the most competitive strategy, for example, involves burning a ton of resources you would rather not burn, that's a big disadvantage: you either don't want to use that strategy, or if you do, it's much worse for you, and you're in a weaker bargaining position than a value system that doesn't care about that form of waste. Or the relevant strategy might involve doing very costly behaviors according to your values. Maybe the most efficient political regime involves causing a bunch of suffering—for example, workers who are exploited on behalf of that regime—so the people who care about goodness don't want to do that. The hope of the "just steal the strategy" story is that you can both use the most efficient means of grabbing territory and then use the territory for your preferences, with no intrinsic disadvantage for one value system relative to another during competition. The concern is that that's not always true.

Fin: Got it. You use this word "locusts" to describe these kinds of fast-growing groups, kind of in contrast to the good stuff. One reaction to all these worries about locust groups emerging and racing to steal the future is: let's stop them from emerging. Let's all agree right now that that would be really bad and it'll be much easier to prevent it than to try to negotiate with it or race against it. That might in fact be kind of easy. I know what you think about that.

Joe: Yeah. In some sense the traditional response is: if there is some value system such that, if it arose and gained a sufficient amount of power or opportunity to start racing, it would quickly outgrow everyone else and do valueless things by everyone else's lights with the available resources, then you either have to prevent that value system from arising or prevent it from reaching the threshold of power that produces a runaway. How hard that is depends on how frequently it would show up by default and by what mechanisms. To the extent we're in a position, with enough coordination technology, to do the first thing I talked about with respect to negative-sum dynamics—i.e., coordinating to resolve them—it may be that this problem can be solved in a somewhat similar way. If you don't have locusts around yet, or at least not with tons of power, maybe you can coordinate to prevent the creation of value systems of this kind. That could involve quite scary restrictions on certain forms of competition, depending on the empirical situation. I'm actually quite disturbed by the types of restrictions on competition that resolving some of these concerns would involve. Part of why I wanted to explore it in the talk is that there are real trade-offs here, but in principle you could do what you suggested, and I think that's to some extent the default.

Fin: Supposing that fails, the vision I have in my mind is this gray, homogeneous swarm that's eating up all of space and filling it with some kind of smokestack, nightmarish thing. I'm wondering if I'm not being imaginative enough—whether there might be redeeming qualities to whatever it is that, by its nature, wins out in the long run.

Joe: Totally. Yeah, so I have a bit on this in the talk as well. I think at the least we should be imagining that hyper-competitive value systems, if there were a bunch of selection for not just strategies but the actual values themselves, we should imagine that they won't be kind of mindless blight. When people think about locusts—locusts do eat your crops very fast, but they're a little bit myopic: the locusts aren't planning to get the most crops. Locusts are actually dumb and they're not hyper-competitive. Humans are vastly more competitive in the long term than literal physical locusts. Same with cancer. If you think of cancer in the human body as trying to take over the universe or something, this is super bad—it just kills its host and dies. None of these things are building spaceships. You've got to build spaceships if you're going to be an effective locust, and it's got to be really good spaceships.

So at the least, you've got to be super-intelligent. You're going to have to have all these same convergent instrumental cognitive features that we see or imagine in paper-clipping AIs and all sorts of different, very powerful systems. That will include having a sophisticated understanding of the universe and really wild technology. So at the least, this is a world with godlike minds;

they might not be conscious, but they're going to know a bunch of stuff and they're going to have cool tech—very cool tech. It might be gray though. It might not be beautiful tech.

Fin: Yeah, so it might be gray but it probably won't be goo.

Joe: I mean, it might be goo at the level of very small stuff, because a lot of— you know.

Fin: It's just—okay, but there's a point of whatever it is you want to do. Let's say it's reaching another star, or building up some industry or whatever is the thing that gets you power. The most effective way to do that is presumably some kind of complicated hierarchical, clever system. It's not goo, right? The best kind of space program is not just a swarm of homogeneous stuff or machines. It's modules with sub-modules that are produced in factories, which are themselves produced by various factories and tools and so on. Maybe this is the kind of natural thing you should imagine: very complex, even if it's gray and lifeless and bad.

Joe: Yeah, and I think that could well be right. It is in some sense the default. Especially insofar as there are a bunch of different things the system needs to do and contingencies it needs to handle, the tech is going to be quite complicated and modular. How much comfort is that? It might be better than literal paperclip replication, but it's like—you look out at the universe and think, "Wow, those really are some detailed machines." Cool.

Fin: Yeah, yeah, yeah.

Joe: And so that might not itself be that much comfort. I think there's an interestingly different version, though, where we go beyond—I'd be more inclined to ask, who are these godlike minds that are running the show? So insofar as there are superintelligences around that have cracked all of the understanding of the universe and are godlike in their understanding, what's the moral status of those beings? How do I feel about them? That's where my mind would go in looking for comfort in a hyper-competitive scenario.

Hypercompetition, value lock-in, and steering future

Fin: Yeah, maybe to pull on that. There's a question about your meta-ethics, right? You might think that what was good in the first place—the stuff you initially cared about that gets outcompeted—if that stuff is the kind of thing you can realize or that you naturally glom onto if you become really smart, then there's a limit to how far you can diverge from it as the hyper-competitive victor of the universe. By hypothesis, these are incredibly smart beings because they needed to be to get this far, and then it seems like for free you get that they appreciate what's good. There's some confusion there. If, on the other hand, you just think that's not how this stuff works—that you just care about what you care about and there's not much more to it—then it feels like you have less grounds for being worried about what happens to the entire universe. It's a bit hand-wavy, but... yeah, I think.

Joe: I would dispute that at a few different levels. I'm processing this as you're sort of like, well, either moral realism, in which case these hyper-competitive superintelligences will get the right morality and I'm fine, or moral anti-realism, in which case I don't really care about the universe. I would dispute both of those.

On the first point, it's actually not true for this sort of scenario, I think, that you should expect moral truth to be realized by a superintelligence by default. Even if moral truth is such that a superintelligence would realize it, and even if it's motivated to some extent, there's an additional problem: if there's any competitive advantage to weakening that motivation, then with enough selection you'll get beings that are bad. Even if they know—it's kind of a weird case—but this isn't that far from the moral universe many moral realists might inhabit, where there's still a problem if power comes apart from goodness in the relevant sense.

Moral realism, even a motivationally efficacious form, gives you some oomph toward goodness being correlated with power, because certain truths do help with effectiveness. An example: mathematics. Right now, having the right mathematics is highly correlated with power because good math helps science, which helps power. But imagine a domain of math where being wrong about it really helped you become more powerful. Even though selection for general intelligence will tend to make you right about math, if you select hard enough for power, you might successfully select for minds that have weird exceptions for that case, or who have double-thought themselves into it, or who are weak-willed about it—you still get the same problem. I think this holds even if you posit some form of moral realism.

Fin: What goes through my head there is: okay, but surely the hyper-competitive victor of the universe that is initially just selected for seeking power seizes every star, and then there's no more competition. Finally it can kick back, reflect on what's good, and things are okay on this realist view. But the issue you raise is the same: if there's any cost at all to thinking about what's good or bringing it about—it's a bit like an analogy to cancer—within this kind of world, if there's any variation, then you just can't cling onto it.

Joe: I do think the empirics matter here. If there's some competitive process that hasn't brought you to the limiting case—where, for example, you've obliterated your motivation to do what's good because that was competitively helpful—then it could be the case that competition has ended in some relevant sense. If competition has ended, then maybe we only need to worry about however far the selection process went up to that point. If there's still some hope, then we're okay.

It sounds like you're wondering what it actually looks like to end competition. Some people who are most pessimistic about avoiding hyper-competitive worlds think competition is deeply endemic to any level of abstraction that allows variation and selection. It's not just "prevent a few wars"; the world is a teeming set of interlocking patterns, and you'd need a kind of fractal totalitarianism to shut down all forms of competition. Obviously we don't need to shut down all forms of competition—depending on the threat model, it might be fine to allow many kinds of

competition and only limit a few, which is what we do in our current economy. So there's a question of how far the limiting case needs to go and what control you can exert.

You could imagine reaching the end of the universe and having a being that's "done," but it still has to govern a civilization where there are all sorts of selection processes and instabilities. The question is how you actually eliminate the relevant forms of selection for the most power-seeking or powerful patterns in that world.

Fin: And one thing you raise is a possibility that competition ends. I mean, it will eventually, right? But there's this risky hope that you might be able to curtail these competitive dynamics indefinitely and just avoid the process running long enough to steamroll what people wanted. A version of that is boot-stomping totalitarianism—an über-surveillance state—and that's maybe even bleaker. But maybe a version of this is super-effective international agreements or giant grand bargains that are just indefinitely enforced, where everyone gets their slice of the future and everyone's making deals and compromises where they all view each other as better off. Maybe that's not, by anyone's lights, the best future, but by everyone's lights it's pretty good. That sounds kind of likely.

Joe: I think if we assume the ability to make these sorts of lasting, stable commitments— which I think is not at all obvious from a technical perspective—you have to really imagine a sort of endemic assumption in a lot of futurism that there's this lock-in potential that then fully stabilizes even a very changing world and has the foresight to do that. That's a very substantive empirical assumption, especially if we have not yet reached the limits of technology.

It's often assumed that that happens at the end of some limiting process of technological development, which isn't necessarily the point at which people are imagining these treaties being signed. But also, if you assume the ability to do that and you assume rationality on the part of all the actors and that it's still a multilateral situation by the time you've reached this lock-in capacity, then in some sense it will follow from your conception of rationality and the options available that actors will engage in some form of coordination to avoid negative-sum dynamics and lock in some bargain. That actually just falls out theoretically of the way you set up the problem. The question is whether the problem has that actual shape—has that degree of rationality and those options available. But yes, if it did, then you get a grand bargain and it might not be so bad.

Fin: It's reasonable. Maybe there's a more realistic version of a bargain where some actors just won't go along with anything. So it's not a fully voluntary deal or a contract—it's more like what already happens in international politics: certain actors throw their weight around when they need to and don't capitulate to actors like North Korea or other rogue states, but otherwise they're in the business of trying to make deals rather than tread on each other. You could be in a position where there's an obvious locust actor that's not part of the deal. Everyone else agrees, for the greater good, to curtail them or try to destroy them. So it's not friendly or mutually

agreeable, but a coalition of the non-locusts forms and things end up such that there isn't a process that just runs into the limit. Again, I don't know how plausible that is.

Joe: Yes, that seems plausible within the theoretical setup we're imagining. I notice I have hesitation around this relationship with "locusts." I don't like the term I introduced; it has a kind of dismissiveness and assumes an enemy quality, like a blight or pest. I feel uncomfortable about that us/them framing. I want our grand bargains, wherever possible, to be inclusive of all actors—including those who might, from someone's perspective, be understood as locusts.

The term "locust" is underdefined with respect to the values at stake. In some sense it's an umbrella term for agents that place intrinsic value on instrumental goods in ways that put you in a worse bargaining position if you're trying to pursue the most instrumentally valuable strategy that would otherwise be wasteful. There are many versions of that. It's not necessarily that someone is a total enemy—there will be people more along that spectrum than others, people who are more comfortable with conflict or who prefer different trade-offs. So I want to be careful about the notion that there is a civilizational enemy, the locust, that we must coordinate to prevent from ever arising, even though I acknowledge that's the sort of framework I used in the talk. I notice I feel uncomfortable about the sort of us/them vibe it's creating.

Fin: Yeah, no, that's totally useful to hear. And there is an open question around — you're talking about this limiting process. Is it actually so terrible? The word "locus" is like baking an answer to that question, maybe prematurely. But it's a question that matters, right? Because if you agree it's terrible, then you want to do something about it. And then you're raising the possibility that although I totally agree it would be lovely if we could make some bargain, some mutually agreeable, non-coerced deal, you're suggesting that with respect to some of these actors that would otherwise just inherit the future, other actors could just feel in a very weak bargaining position, so they might prefer some other course of action. And you might just have to confront that possibility.

Joe: I want us to at least grapple with this possibility as a type of dynamic. And it is possible that this goes quite hard and we end up feeling like, wow, actually good features really require a substantial amount of constraint on the forms of competition that civilization can let run wild. You actually need to be quite careful in channeling competitive dynamics and selection and evolution in the right ways, or you're going to quickly de-correlate with the stuff that you want to keep around.

Fin: As a way of wrapping up, one specific way you might worry about this is you might just really care about distinctively human values, or literally humans themselves sticking around as long as possible or forever. Is that what you're saying, or is that a reasonable view to have, which is: what we all have in common with one another is that we're humans, members of the same species at the same point in time in cultural history. So we should agree that we want to maintain this thing — the biological features, facts about our moment in time as compared to 500 years ago. How open to change should we be is maybe one way of putting this.

Joe: Yeah, so that's very much not what I'm saying. I am not — I want art, and this is part of, to some extent, one of my concerns with arguments for a willingness to curtail certain kinds of competitions. I worry that they'll lead to a kind of lock-in of a really stasis-heavy type of civilization. This is one of the things that's bad about shutting down forms of competition: you end up with the status quo stagnating and sticking around and not being subject to any kind of pressure or accountability. And so I think we want the flower of goodness and joy and beauty and the stuff that we ultimately care about as humans to bloom fully. I think that blooming does not necessarily take the form that we've seen thus far — in fact, I doubt that it does. I want us to be open to quite deep forms of change.

Obviously, moral progress needs to continue. We're very clearly far from the actual final form of human morality — if anything, because our morality is clearly incoherent and not well applied. There's tons of moral progress we need to make. And we just need to learn radically more about the universe and about our place in it and gain wisdom of all sorts. So I think we need to really deeply grow up as a civilization. Very importantly, that means not locking in our current cultural moment, not locking in even, in my opinion, our biological capacities. I want us to respect the human lineage, and I'm not arguing for abandonment or active destruction of the strands of value and history that are already operative within our civilization. But I think what we want is for those to continue to grow and flower in an organic, ongoing-change kind of way.

So it's a very important part of my picture, but it is interestingly in tension with this competitive stuff. Part of what we have trusted thus far in the ongoing process of human development and wisdom is these bottom-up organic competitive selection processes. So you do need to keep those around and allow for ongoing forms of change. You can't just say, "no more competition" — that's just lock-in. You really don't want that. You want to allow for certain forms of change and progress in learning, and you want to do that in a genuinely open way where you don't know what you conclude. You need to be ready to learn new stuff that you're uncomfortable with learning, stuff you didn't get to check ahead of time to make sure you were comfortable with. That's a real balance and part of what makes this problem difficult.

Fin: One thing that strikes me is there is a kind of axis of disagreement, which is just how quickly to kind of welcome technological change and how much to accelerate or decelerate or pause or whatever. There's a seemingly different axis here, which is how much to cling to whatever we've got going on right now. We can agree this isn't so bad, and maybe from here on out it's kind of going to get eroded. So let's just imprint what we care about right now onto the future versus openness to progress and maybe even trying to welcome or accelerate the kind of reflective deliberative processes that seem kind of good, seem like they're going to point in better directions, appreciating how far we might be from how differently better things could get. That's an axis that kind of feels real to me and could end up mattering just as much. But maybe it's worth kind of thinking about what stance to be taking there, and it could be the one that points towards some more openness, some more encouragement of those processes, than we might get by default through standard politics.

Joe: Yeah, you know, I think the answer is in the middle. It's importantly in the middle. We don't want to just exert no control. I mean, I wrote this whole series, Otherness and Control in the Age of AGI, about, in some sense, how much to try to steer the future versus kind of letting the future go unsteered. These sort of yin and yang balance each other.

Fin: One hand on the steering wheel.

Joe: And I think that there is no royal road. I think people have different kinds of vibes. Some people are very yang and some people are very yin. And I think articulating the right structured middle ground is really key. I don't think we as a civilization have a settled answer; to some extent our answer is just the collective expression of all of our institutions at the moment. I certainly don't think we can all agree it's pretty good now.
However, it does look like you've missed a few of the finer points that I wrote in this doc.
Fin: Right?

Joe: The actual world is obviously flagrantly substandard by our own moral standards. So now it's a different question: what would happen if we kept our current morals fixed? But our current morals are also incoherent — we don't have a settled ethic.

Fin: I should quickly say, when I said we have it good or something like that, you're right—that's not true. Maybe it's more like we can broadly agree on what "good enough" looks like, even if we're not there yet. You can see a line to it.

Joe: You could, for example, imagine our current civilization but without cancer.

Fin: Right.

Joe: Let's imagine our current civilization but with everyone having some basic standard of living — you can play that game and get a gradient towards reasonable consensus on many improvements. That's a very minimum bar for imagining what a good future could be. The degree of consensus will start to fray, maybe faster than one hopes. We should have vastly higher aspirations than that.

Fin: Joe Carlsmith, thank you very much.

Joe: Thank you so much for having me.