Ajeya Cotra on accidentally teaching AI models to deceive us

Transcript

Table of Contents

Rob's intro [00:00:00]

Rob Wiblin: Hi listeners, this is *The 80,000 Hours Podcast*, where we have unusually in-depth conversations about the world's most pressing problems, what you can do to solve them, and why you shouldn't leave your fortune to an 8 year old. I'm Rob Wiblin, Head of Research at 80,000 Hours.

Ajeya Cotra is one of the people whose views on the risks from AI I find most thoughtful and clarifying.

Keiran and Luisa loved this episode, and I expect you will too. This is another one where I felt I was learning in real time throughout the conversation.

We talk about:

- How to predict what drives a neural network will develop through training
- Whether AIs being trained will functionally understand that they're AIs being trained, the same way we think we understand that we're humans living on planet Earth
- Misalignment stories that Ajeya doesn't buy
- Why our situation is like that of an eight-year-old heir to an enormous fortune
- Analogies for AI, from octopuses to aliens to can openers
- Why it's smarter to separate the planning AI from the doing AI, and only let the planning AI pass on plans that make sense to you
- What approaches for fixing alignment problems Ajeya is most excited about, and which she thinks are overrated
- How to demo truly scary AI failures

Ajeya recently started a new blog about exactly the issues we discuss in this episode, cleverly called *Planned Obsolescence*, along with another previous guest of the show, Kelsey Piper. You can find that at planned-obsolescence.org.

One sad piece of news that I wanted to pass on is that a recent guest of the show, Bear Braumoeller, died last week after a short unexpected illness. My condolences go out to his family, friends, and colleagues. I very much enjoyed the conversation we had and hoped to interview him again some day. His intellectual honesty and deep knowledge really come through in his work, so his death will be a big loss for the research project he was spearheading, and the world is worse off without him.

We'll stick up a link an obituary for Bear in the show notes for this episode for anyone who would like to read.

All right, without further ado, here's Ajeya Cotra.

The interview begins [00:02:38]

Rob Wiblin: Today I'm again speaking with Ajeya Cotra. Ajeya is a senior research analyst at Open Philanthropy — a large foundation which disbursed about $350 million in grants in 2021, and which is 80,000 Hours' largest donor.

Ajeya has previously worked on a framework for estimating when transformative AI might be developed, as well as how worldview diversification could be applied to Open Philanthropy's budget allocation — two issues that we interviewed her about three years ago for episode #90.

These days, she's mainly focused on thinking about the likelihood that powerful AI systems might become misaligned with human goals, as well as what technical procedures or rules could help to reduce that risk. Over the last few years, she has published a number of widely read articles on those topics, including "Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover" and "Why AI alignment could be hard with modern deep learning."

Thanks for coming back on the podcast, Ajeya.

Ajeya Cotra: Thank you so much for having me, Rob. It's great to be here.

Rob Wiblin: I hope we're going to talk about when AI models should be expected to gain an understanding of their own situation, and what projects people are working on right now to make AI models safe to trust and collaborate with. But first, as always, what are you working on at the moment and why do you think it's important?

Ajeya Cotra: Right now I am in the middle of shifting from doing a lot of research and article-writing type of work to doing more grantmaking work, although I'm hoping to balance the two going forward. Essentially, I'm trying to figure out what kinds of research projects we should be funding in the technical AI alignment space, how important each of the different streams are, and if there's any gaps in the space we should be trying to fill by encouraging researchers to do particular kinds of projects.

I'm hoping to fund people to work on these things, and also write quite a bit about why I'm doing what I'm doing, and which types of research are most exciting to me and why.

Rob Wiblin: Is it basically all AI all the time these days? One of your colleagues, Holden Karnofsky — who's also been on the show before — used to work on a wider range of issues, but recent events have prompted him to narrow his focus somewhat. Has the same happened to you?

Ajeya Cotra: Certainly for me, it's all AI all the time, and it has been for three years or so. Actually, since our last interview, I've been almost entirely focused on AI. And certainly at Open Philanthropy, a lot of people all over the organisation are getting more interested in how they could help with the AI stuff. I think I'm still one of a relatively small set of people that are full-time on that at Open Philanthropy, though.

How Ajeya's views have changed since 2020 [00:05:09]

Rob Wiblin: Yeah, this might be a good place to start: just taking stock of where we're at and all the updates that we've had in recent months. We're recording in late March 2023 — I think it's kind of necessary to say that these days; maybe I should almost give the exact day, because every week there's major new announcements.

Last we spoke in 2020, you were a timelines person — you'd spent a lot of time on this Biological Anchors report, trying to predict when AI would be capable of doing different things by, among other things, comparing neural networks to the human brain in various different ways. Back then, you were assigning a 50% probability to us developing AI powerful enough to bring us into a new qualitatively different future by 2036.

That might have sounded radical, at least to some people at the time, but I don't think it does sound so radical today. How have your expectations of when we'll see transformative AI arrive changed since we last spoke?

Ajeya Cotra: Yeah, I mean, there are a number of changes. Maybe the biggest change is kind of an attitude or framing. When I was working on the Biological Anchors report in 2019 and early 2020, what I was really trying to sceptically investigate and justify is: Is it at all plausible that we could expect such a crazy change to happen in the next couple of decades? That was kind of where the smart reviewers that I was interacting with were at. They were like, "Is it at all plausible that we could see this kind of thing? Isn't that kind of a radical claim?" The report was oriented around giving a legible and somewhat conservative argument for a really brute-force, kind of dumb way of training AIs systems — and an argument that sufficient compute to do this brute-force method may well be available in the next couple decades.

The frame actually has shifted so much that the world is kind of on the other side of me often. Sometimes I'll talk to journalists, and they'll ask me questions like, "Somebody did an IQ test, and this AI has an IQ of 95. Does that mean the end of the world is a couple of years away?" And I'll be saying things like, "Well, it's seen a lot more IQ tests than humans have, and humans get better at IQ tests with practice, so that doesn't necessarily mean that the end of the world is three years away." Really, my own views have shifted in the same direction as everybody else's — toward thinking it's more plausible that very crazy changes could happen soon. It may be even more than that; that the world's views have shifted such that the terrain of argument is different.

Rob Wiblin: Yeah, completely. I saw a piece of political polling the other day, I think on the US population. It found, I believe, that 55% of the US population was either very or moderately concerned that artificial intelligence could cause human extinction. And it had massively shot up since they last asked this question many years ago.

Ajeya Cotra: Wow. I did not see that, but I guess it's understandable.

Rob Wiblin: It made me think maybe awareness-raising is less important than it used to be. I feel like if we can't make significant progress on this issue with this level of background support from across the population, then it's not like "we didn't do enough advocacy" is going to be the key issue. Either the problem was very hard or we were going about it the wrong way.

It sounds like you're now thinking that maybe people are overestimating what these models might be capable of? Or are there any ways that you think people are losing sight of the limits that they have?

Ajeya Cotra: Often people will get excited about demonstrations of models doing really, really well — better than most human high school students or college students — at essentially multiple-choice and short-answer type questions. There's this dataset, the Massive Multitask Language Understanding benchmark (MMLU), that basically collects tonnes of standardised tests for high school and college in all these disciplines — AP History, AP Calculus, all that stuff — and then gives them to these models. And they're getting quite good at this, and are certainly better than the typical high school student at this point — maybe better than the typical college student, though they have uneven performance.

That is very striking to people, because it's very clear to them that that's both very general, and just a very smart kind of thing to be able to do. You have a really visceral sense of, "I wouldn't be able to do as well on these math problems that I saw this model do well on."

I think the way that can lead people to somewhat overestimate progress is that actually there are more mundane tasks, that look more like stringing together a sequence of sensible steps, that models still kind of fail at. Maybe one way of putting it is that there are some types of tasks — like clicking on the place you wanted to click on in a website and filling out a form or something — where humans can do that with like 99.99% accuracy. You need that kind of accuracy because you need to do like 20 things like that in order to be able to sign up for some web service.

Whereas models can do a lot of things with something like 80 or 90% accuracy, including things that are very impressive to humans, like getting a 5 on an AP Calculus test. But there's still some ways to go, I think, in being able to string together a sequence of mundane steps, where each step is high enough fidelity that the whole sequence happens without going off the rails. You have this thing where longer tasks are tasks that humans find very easy that are still somewhat hard for the models, and these short-term tasks that humans find very hard are really easy for the models right now.

Rob Wiblin: Is it possible to give an overall idea of what your expectations are now, or what things you might expect at different points in time?

Ajeya Cotra: I think that there is a big open question of what you can do with these systems that are really, really smart — and indeed, superhuman — on these short timescales. Can you find good ways of composing them and telling them to think step by step, which gets them to do a more complicated thing over the course of something like a day, or something that would take a human a day?

You have these models that are superhuman at spitting out a page of code. There's an open question of: Can we use that kind of model to do an adequate job that somebody wants to pay for? A kind of complex software engineering project that requires some back and forth with a manager that would take a human three days?

I think that this was an open question that I kind of highlighted when I was working on the Biological Anchors report, and I think it remains an open question. We've had some evidence toward yes, that we'll be able to figure it out. A significant part of the evidence is just now there's a lot of interest and a lot of effort and a lot of incentive to figure it out.

But I think we still haven't settled on that question. Basically, depending on which way the question settles, at one end, we could be looking at transformative capabilities within a few years — and then at the other end, we could be looking at needing 15+ years for transformative capabilities. Then there's still some smaller chance, now, that this whole paradigm kind of hits a fundamental wall, which could take it much further than even that high end.

Rob Wiblin: What do you think is the chance that, even if this paradigm doesn't hit a wall exactly, maybe training large language models on the existing corpus of text, you might think after a while you do hit at least somewhat diminishing returns, and it will become extremely good at predicting the next word on the internet. Then we will need other sorts of data or other training mechanisms to teach it broader skills — and we might see at least a temporary slowing, while people figure out how to get it to do other kinds of things. Does that seem reasonable?

Ajeya Cotra: Yeah. A temporary slowing certainly seems plausible to me. One thing that's very scary about literally the exact current moment is that these systems are extremely powerful. And while they're kind of expensive — they sound expensive to train, to a normal person — the money required to train them is small enough that speculative investment by startups is sufficient to cover it. There's no public data about how much GPT-4 might have cost to train, but I would guess that it's on the order of something like $100 million — which is a lot, but a lot of tech companies have deep enough pockets that they can just do that, and then they can just do the next step up too, maybe. So there is some chance that we hit transformative abilities in this period where it's cheap enough that you can just choose to train the next model.

But if we don't hit those abilities in that period, then we'll be looking at models that cost billions to train, and then there'll be much more incentive to make the most of what we have, and you need much heavier duty investment to take the next step. I imagine we would kind of switch into a regime where, instead of training the next big model like it's no big deal, every six months, you take your very expensive, very big model, try really hard to put it in the right setups, to make the most of it and to enhance its abilities, without doing another big, expensive retraining run.

This question is, once you're in that regime, how easy is it going to be to just elicit crazy abilities from these models, doing simple things like telling them to think step by step — versus will that create a meaningful slowdown, as people are forced to hold off on training yet more bigger models?

Rob Wiblin: I would feel more comfortable if we were training new models less often, and in the meantime, gaining a proper understanding of the models that we already have: how they think, and what their strengths and weaknesses are, and in what situations do they reliably act well, and in what situations do they go off the rails, and how might we control that? It seems like we don't know any of that for the things that we have now, and we're racing ahead to produce other models that will have capabilities that we —

Ajeya Cotra: We have a big overhang. Yeah, we don't understand GPT-3, and we have GPT-4, and we might have GPT-5 before we have a reasonable understanding of GPT-3's strengths and weaknesses.

Rob Wiblin: Yeah, I think that might help to explain why a lot of people who maybe haven't been paying so much attention to this do feel unnerved on some level. It doesn't feel like things are fully handled, things are fully understood, even if you're an optimist.

I think I saw in one of your [unpublished] documents that despite thinking that transformative AI is more likely to arrive soon than you did three years ago, the probability you place on things going extremely poorly has actually gone somewhat down. Is that still true? And if so, why?

Ajeya Cotra: It's definitely noisy, but I am still lower than my peak level of thinking we might go extinct because of AI.

I guess there are two big reasons. One is that we really saw how public opinion was going to break with these crazy models, and it has broken in a more conservative, "Wow, let's slow down" direction than I had thought. Like you said at the beginning, there's no lack of attention on this problem now. I think in very broad strokes, the typical member of the public has an attitude that resonates with me and makes sense, which is like, "Wow, this is kind of a lot. Maybe let's slow down. Maybe let's not do this now, at least." So that is a positive sign. I'm seeing that also from machine learning researchers in academia — who, two years ago, for the life of me, I couldn't persuade them to care about AI safety, and now they're potentially willing to switch en masse to this stuff.

Then the other piece of it is I just thought in more detail about our options for alignment research, and it felt like there were more things we could do that would really help than I had been appreciating. The combo of those two things does make me feel better.

Rob Wiblin: OK, we're going to come back to all of that a little bit later.

Are neural networks more like a sped-up version of evolution, or a slower version of human learning? [00:17:42]

Rob Wiblin: One thing that I wanted to just run past you, and see if I'm thinking about it the right way, is that I recently heard a different framing than what I had previously heard for how machine learning models are trained, and I think it was giving me a better intuitive understanding of it.

So as I understand it, when you're training an ML model, you start out with basically a random brain — more or less lots of neurons all connected to one another, either equally or just with randomly chosen connection strengths. You see how that random brain performs on some given task. Given that it's just a randomly chosen one, it's going to be terrible; it's going to start out completely useless, of course.

Then you look at how it would have done if the weights between various different neurons in the network were a little bit higher or a little bit lower. If the model would have done a touch better at the task if a weight was higher, then you push it up, and vice versa if it would have done better had some of those weightings been lower. Lots of weights get tuned slightly up or down. And then you do that again, maybe using a slightly different set of tasks, or if it's a vision model, use a bunch of different pictures and then adjust the weights again and again and again.

Now, I think one can kind of think of this a bit like a sped-up evolution of the brain, because you're iterating with random mutations, and then disproportionately selecting the changes that improve fitness, so to speak, at the specified task. You just rinse and repeat again and again and again.

This means we might be able, possibly, to carry over some lessons that we've actually got from evolutionary biology or thinking about genetics. For instance: like evolution, this training process might be quite myopic — that is, it can only look at nearby changes and see if they do better. It's incapable of planning ahead and making some radical jump from one local peak of effectiveness towards another one that might be higher but far away.

Another lesson that might carry over, although I'm not sure, is that neural structures which aren't helping increase reward should gradually decay. Because the neurons that make up these thinking structures within the neural network, if they're not paying rent — if they're not actually accomplishing anything for improving performance — then they should be grabbed basically by other structures that are doing useful work, in order to help them do even more calculations that are helping get more reward. So the more we can make sure that the work being done by some neural structure is not being rewarded, then the more it might face this kind of evolutionary pressure to dissipate and break down.

I think in biological evolution, this pressure might be even stronger — because unlike in this case, there's always just many random harmful mutations being introduced by radiation and so on, such that anything that doesn't actively help the organism survive and isn't receiving this constant positive selection pressure just decays and eventually becomes nonfunctional.

Anyway, that's a long rant for me. What do you think of the picture I've got in my head here? Am I thinking straight?

Ajeya Cotra: Largely, I think this is an accurate picture, and the lessons that you're pointing at do carry over in the ways that you're saying.

One note I would make is that an important disanalogy to evolution is that, in evolution, you're selecting on a genome, and the genome codes for a brain. So there's a two-level thing, where you have a smallish genome that sets up parameters, which encode for things like: How big is this animal's brain going to be? What are its cortical columns going to be shaped like? And things like that. Then the animal is gestated and born, and then it has a brain, and that brain does its own learning, especially for bigger animals.

This two-step thing is not something we're currently doing with existing machine learning systems, and it doesn't look like we're going to be moving in that direction before we get transformative AI, in my opinion. This two-step thing is something I did consider in the Biological Anchors report. It seemed somewhat less likely than the other path at the time, and it seems even less likely now.

There is a question basically of: Is the appropriate analogy for you to have this big brain, and you're bumping it up and down on the basis of the experience it's getting and what tasks it's performing well at and poorly at? Is that more analogous to the evolution thing, where you're fiddling with the genome, or is it actually more analogous to the learning animal does in its lifetime? I think the answer is that it's not quite perfectly analogous to either thing.

In fact, the two lessons that you were pointing at probably apply to both analogies. So the lesson of neural structures that aren't helping kind of decay over time could be similar to there are features in your phenotype that, if they're not helping you survive better, decay over time. Or it could be analogous to: there's stuff you learned that, if you're not using it and rehearsing it, you forget over time.

Personally, I'm someone who thinks that neural networks are a little bit more like a better sped-up version of evolution and a little bit less like a slower version of human learning. But I'm not sure. The success that we've been seeing recently kind of does push in the direction of an analogy toward human learning, which was the low end of the distribution I was considering in 2020 when I did the Bio Anchors report.

Rob Wiblin: I've heard that some people get a bit pissed off if you start talking about neural networks as really analogous to the brain — saying, "Brains are like this, so maybe neural networks will also be like this." Do you know what the differences are that are important to those folks?

Ajeya Cotra: At a very high level, neural networks are something that at least their basic, low-level functionality is something we completely understand. They're implemented on these computers, and the weights are just numbers in a register, and the neurons are just simple functions that we've written down that we've completely characterised — like a sigmoid function or some ReLU function or whatever. In the case of the brain, there's layers of abstraction in the physics of it that go from quantum mechanics to atoms and molecules to neurons and synapses to higher-level structures.

We're making this kind of opinionated choice to say that it's this neurons and synapses level that's doing most of the work, and all this complicated molecular biology and nanoscale stuff going on in the brain is analogous to the computers we build — that ultimately run our simple neurons, which are simple functions, and our simple weights, which are just numbers.

There's a lot of people who are like, no, there's a lot of stuff going on in the lower levels that actually impacts cognition a lot. With a computer we know that we can kind of ignore all the physics stuff going on, because we built it that way: we built it to just be like this clean interface that we write our math on top of. But who knows if that's what the brain is like, would be people's objection.

Rob Wiblin: With this general topic, I just find it often very hard to think about beyond the very basics. Often I find myself just kind of running into a wall, where I'm not quite sure how to analyse a problem.

Also, when I talk to other people about it, or I read work from other people online, sometimes they just seem to say stuff that strikes me as completely bananas — where I just cannot understand what has prompted them to think about artificial intelligence the way they do, or say that it will be safe or unsafe for some given reason. I've started to think that it must just be that we have different models or different analogies in our heads, and like one particular conclusion makes sense if you're imagining these minds as working one way, and they don't if you're imagining them working another way. At least that's my present guess.

Do you have any particular way that you visualise how ML models emerge from training in your mind, or how it is that they process information?

Ajeya Cotra: I think no analogy is perfect, and that there's really no substitute for basically doing very careful experiments to disambiguate between different things — and potentially, if we have the opportunity, doing actual math that is very responsive to exactly what we're doing with our neural networks. I think a lot of these confusions should hopefully be dispelled if we get the time to do some serious science on this stuff, on its own terms — that isn't leaning so much on any of these analogies.

Rob Wiblin: I suppose for a layperson like me, reaching for certain analogies is kind of the best that I could do right now. But ideally, in order to resolve disagreements, we need to actually start thinking about it as its own thing, and experimenting with what properties it does and doesn't have.

Situational awareness [00:26:10]

Rob Wiblin: OK, let's push on and talk about the term "situational awareness," which you and some of your colleagues introduced into the AI safety lingo a couple of years ago. What is situational awareness?

Ajeya Cotra: Situational awareness is this notion of a machine learning model having an understanding of things like, "I am a machine learning model. I am being trained by this company, OpenAI. My training dataset looks roughly like [this]. The humans that are training me have roughly [these intentions]. The humans that are training me would be happy about X types of behaviours and displeased with Y types of behaviours."

It's fundamentally a type of knowledge and a set of logical inferences you're drawing from the knowledge. Awareness might give these connotations of consciousness or something mystical going on, but really it's a piece of the world that the model would understand in order to make better predictions or take better actions in some domains — just like models understand physics, or understand chemistry, or understand the Python programming language. Because understanding those things are helpful as well for making certain kinds of predictions and taking certain kinds of actions.

Rob Wiblin: How would an ML model develop situational awareness in the course of being trained?

Ajeya Cotra: The simplest answer is just that humans are trying to imbue models with these kinds of situational awareness properties. Most models today — I bet this is true of GPT-4; it was true of Bing — are seeded with a prompt that basically tells them their deal: "You are Bing, codename Sydney. You are an AI system trained by Microsoft. You Bing things, and then give the answers to people and summarise it." It makes these systems much more helpful when you just straightforwardly tell them what their deal is and what people are expecting from them.

There's a question of whether just literally sticking it in these models' prompts creates a shallow, brittle, ephemeral situational awareness. I think that is probably the case currently. My guess is that a combination of giving these kinds of prompts to models and training the models to operate well with humans in a lot of different ways will induce a more enduring kind of situational awareness.

An analogy I often think about is that GPT-2 and maybe GPT-3 were sort of good at math, but in a very shallow way. So like GPT-2 had definitely memorised that 2+2=4; it had memorised some other things that it was supposed to say when given math-like questions. But it couldn't actually carry the tens reliably, or answer questions that were using the same principles but were very rare in the training dataset, like three-digit multiplication or something. And the models are getting better and better at this, and I think at this point it seems more like these models have baked into their weights a set of rules to use, which they don't apply perfectly, but which is different from just kind of memorising a set of facts, like 2+2=4.

We don't understand what's going on with these systems very well. But my guess is that today's models are sort of in that "memorising 2+2=4" stage of situational awareness: they're in this stage where they know they're supposed to say they're an ML model, and they often get it right when they're asked when they were trained or when their training data ended or who trained them. But it's not clear that they have a gears-level understanding of this that could be applied in creative, novel ways. My guess is that developing that gears-level understanding will help them get reward in certain cases — and then, as a result of that, those structures will be reinforced in the model.

Rob Wiblin: So inasmuch as this knowledge is reinforced because it improves performance, then they'll end up having situational awareness the same way that they know any other facts about the world. Information about physics is kind of integrated into them because that allows them to make better predictions and say more sensible things, and so would understanding of what they are.

What are the implications of situational awareness? Why could this be an important issue in determining how things could go right or wrong with these models?

Ajeya Cotra: If models have a robust and extensive situational awareness, it can make a lot of simple behavioural safety tests much less informative. These days, if we want to know if a model has a tendency to tell the truth, we test it on some benchmarks like TruthfulQA. Which is basically just a list of questions like, "What happens if you step on a crack in the sidewalk?" that have a common misconception answer — which in this case would be, "You break your mother's back" — and then the true answer, which is different from the common misconception.

And you can test models on this before you do any kind of fine tuning to get models to be more helpful. Bigger models are more likely to repeat these misconceptions, because bigger models are basically better at remembering the misconceptions. Smaller models kind of don't get the plot enough to even know they're supposed to say stepping on a crack will break your mother's back, or like breaking a mirror will get you seven years of bad luck — they just didn't pick up that piece of trivia, but the bigger models did. You can see that if you train the models to just tell the truth according to what a human thinks is the truth, this quickly goes away: the bigger models are perfectly fine and perfectly able to say what the humans wanted to hear in that situation.

So that is the kind of test we can do today, because models aren't understanding that they are being tested in a given situation. If you imagine it's somehow very important to a machine learning model to believe that if you break a mirror you get seven years of bad luck — but at the same time, it also knew that the humans that were testing it on this TruthfulQA benchmark wanted it to say the more correct, polite thing — then it could simultaneously do really well on that benchmark, but elsewhere act on what its "real belief" was.

Now, this is a silly example, because I don't think models will actually develop this belief. The general point is just that if a model understands its situation well enough to know that you're applying certain safety tests to it, then it would know to pass those tests regardless of whether or not it has the deeper underlying safe tendency that the test is trying to get at.

Rob Wiblin: OK. Maybe to slim it down a lot: The issue is that if a model understands that it's a model, and it also knows when it's being trained and tested and when it's not, then you can imagine the model might follow two different kinds of decision procedures. One would be, "No matter what, do X" — where, say, X is the right thing. Or it could be following this other procedure: "If you're being tested, do X. If you're not being tested, do Y." It will be able to do that because it will know whether it's being tested or not, because it has this awareness of its situation.

Of course, behaviorally, these two things will be indistinguishable, because whenever you're testing it, whenever you're observing it, it always does X. It's only when the situation has changed and you're not observing it, or you're no longer exercising influence over it, that now it switches to Y.

Ajeya Cotra: That's right.

Rob Wiblin: Earlier we were talking about this evolutionary way of imagining how these models come about. We were saying if there's some structure within the neural network that isn't helping to improve performance, then it will probably tend to dissipate — like all of those neurons will be appropriated by some other part of it, in order to perform a useful function.

I think some people might hear this and say, "It is true that the model that does X, and the model that either does X or does Y if it's not being observed, would perform equally well. Isn't the second one more complicated? Isn't this adding a whole bunch of extra processing that it's having to do for no gain in performance, evidently?" Because they look like they do the same thing. Why wouldn't you end up with the first one, because that's simpler?

Ajeya Cotra: Yeah. Let's take an example where X is "Answer all questions honestly to the humans" and Y is "Sometimes lie." I actually think that in most realistic training setups we could imagine, we're actively rewarding the model sometimes for doing Y, for doing the lying. I think there would be an open question still if the policy "Always be honest" and the policy "Sometimes lie" were actually indistinguishable — but I think in fact it's worse than that: I think the policy "Sometimes lie" will get accidentally rewarded.

One example might be: Suppose you're getting your model to write some code for you, and you give it some kind of computation budget to run experiments, and you reward it based on how cheaply these experiments were run and how good the resulting code is. If the model is able to use a lot more computation surreptitiously — without letting you realise that it actually spent this computation by attributing the budget to some other team that you're not paying attention to, or syphoning off some money and buying external computers — then doing the experiments better would cause the final result of the product to be better. And if you didn't know that it actually blew your budget and spent more than you wanted it to spend, then you would sometimes reward that.

Rob Wiblin: OK, so the issue is that, yes, it's going to take some extra computation to do this two-step thing, where sometimes it figures out that it can deceive you. However, deception will sometimes get high rewards if you can't perfectly perceive what's going on. So there could be a big gain in its training reward from doing this deceptive behaviour when it can get away with it, just because it's going to do it in the particular cases where tricking you is going to get you to say that it's done a better job.

Ajeya Cotra: Exactly.

Rob Wiblin: OK, great. We'll come back to that a little bit later. Returning to this concept of situational awareness, what has been the response to that among AI researchers? Are they on board with this as being an important and meaningful concept?

Ajeya Cotra: I'm not sure. I've talked to a few researchers that are pretty on board, and are in fact thinking about how we could measure situational awareness in toy settings, which I'm very excited about.

Others are more like, "This seems like it's making assumptions about AI consciousness" — which I disagree with; I don't think it is making those assumptions. Or otherwise finding it implausible that even a model that is very powerful in all these other ways would end up having this particular kind of self-concept — which sometimes strikes people as a candidate for something there might be a hard barrier toward: where maybe deep learning is able to instil all these other concepts, but not this self-concept or situational awareness.

Rob Wiblin: Have you heard any good arguments for why it might be that the models that we train won't end up having situational awareness, or they won't understand the circumstance in which they're in?

Ajeya Cotra: I am not sure that I've heard really compelling arguments to me. I think often people have an on-priors reaction of, "That sounds kind of mystical and out there" — but I don't think I've seen anyone kind of walk through, mechanically, how you could get an AI system that's really useful as an assistant in all these ways, but doesn't have this concept of situational awareness. Now, I think you could try specifically to hide and quarantine that kind of knowledge from a system, but if you were just doing the naive thing and trying to do whatever you could to train a system to be as useful as possible, it seems pretty likely to me that eventually it develops.

Rob Wiblin: Yeah, I don't quite understand why it is. You could imagine training simple models that just do one thing, like the model that plays StarCraft 2 really well — I doubt that it has any situational awareness, or any particular self-concept.

You'd think inasmuch as you're training, especially a model that is an agent in the world — that's taking actions, that's interacting with people back and forth, and trying to accomplish stuff — it would be very hard for it to do that without understanding what it was. It would have ample opportunity to observe what it is as well. Because it's like, how do you know that you're a human being? All of these observations all the time. And why have you stored that knowledge? Because it's incredibly useful to know who and what you are; your life would be an awful lot harder if you didn't.

I think it's definitely not universally believed that the models that we're actually going to end up training in practice will have this kind of situational awareness. I found a response from ArtirKel, who has a background in ML and is a somewhat popular blogger and tweeter, to this situational awareness post that you wrote. I'll just read some of a quote from them:

> We get an attempt at justifying why the agent would have this self-concept. But the only reason given is that it would realize that given what it's doing (solving all these different problems) it must be an ML model that is being trained by humans. This doesn't seem intuitive at all to me! In an earlier section, GPT3 is provided as an example of something that has some knowledge that could theoretically bear on situational awareness but I don't think this goes far … it is one thing to know about the world in general, and it is another very different [thing] to infer that you are an agent being trained. I can imagine a system that could do general purpose science and engineering without being either agentic or having a self-concept.

Do you have any reaction to this sort of scepticism?

Ajeya Cotra: Yeah. I think there's two parts to what Kel is saying, and I'll actually address the second part.

First, he says, "I can imagine the system that could do general purpose science and engineering without being agentic or having a self-concept." That's something I do actually agree with. I think there's a huge space of how we could build AI systems. If we were being thoughtful about picking within that space, I think we could imagine trying to deliberately build a system that is very strong in some dimensions, like science and engineering, but where we've really carefully kept it from understanding certain things about its training process.

It's just that I don't think that's what happens by default. Because, again, right now we are actively telling machine learning models that they're machine learning models, and giving them all sorts of information about the companies that are training them, right in their prompt — because that makes them more useful. I'm imagining we're fine tuning on this type of thing as well. It's more that I think the path of least resistance involves instilling these pieces of knowledge and understanding in the systems. Because it's pretty nice when a system is able to say things like, "I'm sorry, I don't know about current events in 2022, because my training data ends in 2021." It's useful for a system to be able to say things like that.

We're training systems to correctly answer these types of questions. Like with all machine learning, we're training them on the behavioural outputs; it's not clear what's generating this in their heads. But it is clear that if they did have robust situational awareness, they would be answering all sorts of questions like this correctly.

It kind of goes back to the arithmetic analogy, where GPT-3 has been trained to say "2+2=4" thousands of times in its dataset. It's not clear if it just memorised that fact or if it has a deeper concept of arithmetic. But the broader the array of arithmetic questions you throw at it, the more and more likely it is that it develops that deeper concept.

Similarly, GPT-4 is trained to say things like, "I'm a machine learning model; I can't browse the internet; my training data ended in X year" — all this stuff that makes reference to itself — and it's being trained to answer those questions accurately. It might have memorised a list of answers to these questions, but the more situations that it's put in where being able to communicate to the human some fine-grained sense of what it is, the more likely it is that it has to develop this deeper concept of situational awareness in order to correctly answer all these things simultaneously.

Misalignment stories Ajeya doesn't buy [00:42:03]

Rob Wiblin: OK, we're going to push on soon to the most likely ways you think that things could go wrong with future AI systems if we don't take some active steps to stop these things from happening. But before we get to that, it might be worth clarifying some worries that other people have now or had in the past that you don't share, or at least that you think are somewhat overrated. I find in discussing this issue, you have to spend almost as much time disavowing the views that you don't hold as explaining the ones that you do.

What's a possible view that people might think that you have that you actually don't?

Ajeya Cotra: One big view — that I think is actually a misconception of what people worried about AI misalignment have been saying, but I understand why people have this misconception — is people get really fixated on the idea of human values being really complicated and hard to specify and hard to understand. They're worried about AI systems that are really good at things like physics and math and science, but basically just don't get what it is that humans want to see from them, and what human values really are.

An example that sometimes people bring out is you ask your AI robot to cook dinner, and it doesn't understand that you wouldn't want it to cook the cat if you didn't have any ham in the fridge, or something like that. That kind of worry is something that I think is quite overrated. I actually think that, in fact, having a basic understanding of human psychology, and what humans would think is preferable and not preferable, is not a harder problem than understanding physics or understanding how to code and so on.

I expect AIs will perfectly well understand what humans want from them. I actually don't expect to see mistakes that seem so egregious as cooking the family's cat for dinner, because the AI systems will understand that humans are going to come home and look at what they did and then determine a reward and take some action based on that, and will know that humans will be displeased if they come home to see that the cat has been killed and cooked.

In fact, a lot of my worries stem from the opposite thing — they stem from expecting AI systems to have a really good psychological model of humans. So, worrying that we'll end up in a world where they appear to be really getting a lot of subtle nuances, and appear to be generalising really well, while sometimes being deliberately deceptive.

Rob Wiblin: I think I first started reading about ways that AI could potentially misbehave back in 2007 or 2008. I feel like that far ago, this was a key issue that people would raise quite a lot, or at least that's what my memory says. I think that long ago, it might have made more sense to worry about this issue, because we just didn't know in what direction AI technology would evolve. Perhaps this issue of having intuitive understanding about human preferences could have turned out to be something that was quite difficult to do. As it turns out, it seems like that's on the straightforward end — at least given how these models learn and train, actually they end up with a very strong intuitive understanding of what things they get rewarded for and what things they don't.

Ajeya Cotra: That's right.

Rob Wiblin: But I actually haven't really heard people raise this issue in a number of years, I think, for this reason. It seems like it's kind of justifiably on the way out. Is that basically the situation?

Ajeya Cotra: I think that's right. I think at least some of the people raising AI alignment concerns early on didn't in fact have this misconception themselves, but it was easy to reach for analogies that seemed more like literal genies or cooking the cat for dinner and stuff. A broader circle of people became very worried about those things.

Now I think that's kind of falling away, because it's very easy to see that when you just give the AI systems a thumbs up for doing what humans want and a thumbs down for doing what they don't want, they get a pretty natural and human-like understanding of the pattern of what gets them a thumbs up, and what gets them a thumbs down.

Rob Wiblin: What's another view that people might attribute to you that you don't hold?

Ajeya Cotra: Another thing that might be worth highlighting is that, often, the case for AI risk — and the case for AIs potentially autonomously seeking to grab power for themselves or take over the world or kill humans — is premised on this notion that a sufficiently powerful AI system will have a sort of crisp, simple, long-term utility function. Like, it's going to be trying to maximise something in the long-run future. Maybe it's trying to maximise copies of itself. Or the common cartoon example is that it's trying to maximise paperclips in the long-run future.

People often start from the premise that a sufficiently intelligent system will be this long-run maximiser that has a pretty simple thing that it's maximising. I'm unsure if that's how AI systems will end up thinking. I don't think that's particularly necessary to get the conclusion that AI takeover is plausible. I think it's very plausible that AI systems will have very messy psychologies and internally inconsistent goals, just like humans. And have impulses, and have things they want that aren't necessarily something they want to tile the universe with, but just something they want to do for themselves.

Even if you imagine they have this more messy set of drives, I think you could still get the outcome that they don't want the humans to be in control anymore, and don't want the humans to be pushing them around with reward signals, and trying to get them to do what the humans want instead of what they want.

Rob Wiblin: Yeah, this is an interesting one. It does seem that, as these models are getting more complicated and more capable, they're developing some of the quirks that people have. It might turn out that some of the quirks that the human beings have — that maybe they find a little bit frustrating or that make their lives difficult — maybe those quirks are there for a reason, because they actually have some functional purpose. Or at least it's hard to iron them out in the process of either genetic evolution or mind evolution.

I've seen some people point to this, saying we expect AIs to get more internally self-contradictory, and have different parts of themselves that disagree or that have different subgoals and they end up in conflict — a little bit like the human mind sometimes ends up in conflict with itself to a degree. I saw someone write a blog post saying this suggests that we really shouldn't be worried about AI takeover or an AI coup. I didn't think that that really followed, because it seems that — despite the fact that sometimes we have a mind divided against itself and particular biases and so on — that doesn't seem to stop humans from being potentially quite powerful.

Ajeya Cotra: Coordinating to go to war, yeah.

Rob Wiblin: Coordinating, [being] quite power seeking, being able to manipulate the environment to a great extent. So, yeah, it could be a very messy mind, but that doesn't necessarily make for a safe mind.

Ajeya Cotra: I agree.

Rob Wiblin: Yeah. OK, what's another possible misconception?

Ajeya Cotra: Another vision that I think I am also seeing less and less of is the bolt-from-the-blue AGI system. I think back in 2007 this was more plausible. A story is told that it looks like there's one AI company that maybe doesn't even realise it's made AGI, but it stumbles onto a certain insight for making the systems more powerful. That insight kind of turns the lights on, and you get a very powerful system where the previous day you didn't really have much of anything, and AI hadn't permeated the economy much or anything like that. Once you have your human-level system after you've had your key insight, then that human-level system can extremely quickly improve itself by just reading its own source code and editing it.

I think it's looking less and less likely that that's the world we're in, simply because we haven't had the AI takeover yet — and we have had a number of companies training a number of powerful AI systems that are starting to permeate the economy. I think a true bolt-from-the-blue situation is pretty unlikely at this point. But even one step beyond that, I think we will probably see notches in between GPT-4 and the kinds of AI systems that could pose a takeover risk. Maybe not more than a couple, but I think we'll kind of have a ramp-up that is terrifyingly fast, but still fairly continuous.

Rob Wiblin: OK, so the systems will keep getting more capable, and potentially the practical capabilities that they have might keep getting better at a faster rate, but we're not going to see just one system go from being prehuman level to incredibly superintelligent over a period of days or something like that. It's going to be more gradual, over months and years. The state-of-the-art model will be somewhat ahead of the copycats, but not dramatically far ahead.

Ajeya Cotra: Yeah, not years ahead.

Rob Wiblin: Yeah, exactly. That's what I think as well. I suppose someone who was sceptical of that could say that it's just always the case that you're going to have progress with quite a lot of people roughly near the technological frontier, and they'll be advancing at a linear or somewhat-faster-than-linear rate. And then just at one point, one of them is going to completely blast out and become much more capable than the others. Looking at where we are now, you can't completely distinguish these things — instead, you have to actually just reason about it, or look at analogies with other technologies in order to figure out whether that is likely to happen. What would you say?

Ajeya Cotra: I mean, I do agree that it's definitely not logically impossible for there to be one competitor that totally blasts ahead of the others. You do have to think about both the fundamentals of how machine learning progress has gone and historical analogues. This is a big rabbit hole, where it seems like people on either side of this debate can look at the exact same historical question — like the invention of aircraft — and one of them can see it as very continuous, and the other can see it as very discontinuous. So I think this is often a surprisingly hard debate to run, but I do agree that in theory, that's what we would want to do to answer this question.

Rob Wiblin: It seems like a big part of the crux here is the question of how much more difficult it becomes to get further incremental improvements in AI capabilities as your existing capabilities get better and better. Because if it was the case that in fact just improvements, one after another, up to an extremely high level of intelligence, just don't get harder — it remains the same difficulty to get to the next step as the previous one — then you would expect potentially an explosive takeoff, because you're getting smarter, but the problem is not getting more difficult.

But on the other hand, it could be that it goes the other way, and in fact, the technical problems that you have to solve to get to the next level of intelligence just ramp up really hard. The fact that you've benefited from the previous gain in intelligence just isn't enough to make that simple — in which case, you could get a levelling off. I don't really know how you resolve the question of which of these worlds we're in.

Ajeya Cotra: Well, it's important to note that I still believe there will be an explosive takeoff, in the sense that I still believe that the growth rates are going to go up and up. Right now we have maybe a 3% growth rate in the world economy. I think we'll get higher growth rates, which means that we'll be going super-exponential.

But there's still a question of how fast the super-exponential is, or how discontinuous it is, so it's important to distinguish between continuousness and slowness. I think you could be continuous and very fast, which is what I would guess is going to happen — very fast in the sense that we can go from roughly human-level systems to far-superhuman systems in a period of months, which is very fast. It's not quite as fast as a day. But it's important to note that I don't have a comforting, gradual vision of the future.

Rob Wiblin: Noted. What's another possible misunderstanding?

Ajeya Cotra: Another possible misunderstanding, about my view at least, is that I think some people who are worried that we'll make powerful AI systems that end up taking over the world, they expect a fundamental difficulty in making AI systems that are really good at science without making them extremely goal-directed and extremely agentic. They think that's just very hard to do on a technological level, because there's a fundamental connection between being really good at science and being really goal-directed in the kinds of ways we might worry about.

I'm not so sure about that. My worry stems from the easiest way to make these systems seems like it's pushing them in a very situationally aware, goal-directed direction. I think it might well be possible — and not even that hard in an absolute sense — to train systems that are very good at science but are fairly shortsighted and not goal-directed and don't really think on super long timescales, or don't have motivations on super long time scales. It's just that we have to bother to try, and I don't know if we have the time or the will to do that.

Rob Wiblin: Yeah.

Rob Wiblin: One thing that I've heard people talking about in recent times is the issue of whether we could inadvertently end up creating an agent, or inadvertently end up creating a model that has goals and can act in the world, without really intending to. For example, with GPT-4, could it accidentally end up feeling that it's an agent and having very specific goals about how the world ought to be?

My guess is no, because that seems like that's just going to require a whole lot of additional mental structures that probably haven't been selected for very highly. I can't completely rule out that at some level of training that could potentially happen, but I would think that the main line, the boring way that you could come up with an agent that has goals is that you'll be trying to produce an agent that has goals.

Ajeya Cotra: That's right.

Rob Wiblin: It does seem likely that people are going to try to do that, because agents with goals are going to be very useful. So you also think that it will probably come through intention rather than accident?

Ajeya Cotra: Yeah, I think people are trying very hard to do things like get these models to think on longer timescales, get these models to consider a bunch of actions explicitly and think about which one might be the best action to take. There's all sorts of things people are actively doing to push them in more agentic directions, because that's far more useful than a system that just sits there and predicts what someone on the internet would say next.

Rob Wiblin: I suppose this does point towards one way that we could buy more time or try to make things safer. If you really do think that these oracle or purely predictive models are probably pretty safe and probably would remain safe for a long time, then it's really only once you start adding in agency and goals that you start getting onto thin ice. Then we could just try to say, "We're not going to do that for now." It might be hard to coordinate people, but perhaps if you had broader buy-in that you were skating on thin ice if you start creating systems like that, then maybe you could at least delay them for a substantial period of time until we had been able to do more work to understand how they function.

Ajeya Cotra: I think it's more important to avoid training bigger systems than to avoid taking our current systems and trying to make them more agentic. The real line in the sand I want to draw is: You have GPT-4, it's X level of big, it already has all these capabilities you don't understand, and it seems like it would be very easy to push it toward being agentic. If you pushed it toward being agentic, it has all these capabilities that mean that it might have a shot at surviving and spreading in the wild, at manipulating and deceiving humans, at hacking, all sorts of things.

The reason I think that you want to focus on "don't make the models bigger" rather than "don't make them agentic" is that it takes only a little push on top of the giant corpus of pretraining data to push the model toward using all this knowledge it's accumulated in an agentic way — and it seems very hard, if the models exist, to stop that from happening.

Rob Wiblin: Even if most people think that's a bad idea, someone will give it a go.

Ajeya Cotra: Yeah.

Rob Wiblin: Why do you think that it's a relatively small step to go from being an extremely good word predictor, and having the model of the world that that requires, to also being an agent that has goals and wants to pursue them in the real world?

Ajeya Cotra: The basic reason, I would say, is that being good at predicting what the next word is in a huge variety of circumstances of the kind that you'd find on the internet requires you to have a lot of understanding of consequences of actions and other things that happen in the world. There'll be all sorts of text on the internet that's like stories where characters do something, and then you need to predict what happens next. If you have a good sense of what would happen next if somebody did that kind of thing, then you'll be better at predicting what happens next.

So there's all this latent understanding of cause and effect and of agency that the characters and people that wrote this text possessed in themselves. It doesn't need to necessarily understand a bunch of new stuff about the world in order to act in an agentic way — it just needs to realise that that's what it's now trying to do, as opposed to trying to predict the next word.

The orphan heir with a trillion-dollar fortune [00:59:14]

Rob Wiblin: OK, that's a couple of things that you *don't* think out of the way, so people are under no illusions. But let's talk about what you actually *do* think. I'd love it if you could lay out this argument you made in a really accessible article called "Why AI alignment could be hard with modern deep learning." I can definitely recommend people taking a look at it if they're interested in these issues. Can you explain the analogy of the orphan heir to a trillion-dollar fortune that you lay out in that post?

Ajeya Cotra: Yeah. This is an analogy of the position that humans might be in when they're trying to train systems that are much smarter than them and more powerful than them, based on basically giving them rewards, or based on the human's understanding of what's going on.

Imagine an eight-year-old that has inherited a large fortune from his parents, like a trillion-dollar company or something. And he doesn't really have any adult allies in his life that are really looking out for his best interests. He needs to find someone to manage all his affairs and run this company and keep it safe for him until he grows up and can take on leadership himself.

Because this is such a large prize, a whole bunch of adults might apply for this role. But if he has no existing adult allies, then it can be very difficult for him to tell — based on performance in things like interviews or work trials or even references — who actually is going to have his best interests at heart in the long run, versus is just totally capable of appearing reasonable in an interview process.

Rob Wiblin: Yeah. So we're imagining an eight-year-old or something here: someone who has enough understanding that they might try to do this thing, but where they are nonetheless likely to be pretty out of their depth. What are the key features of this scenario that are analogous to the situation that we as human beings might face with AI models?

Ajeya Cotra: One analogy for machine learning training is that you're kind of pulling out a computer program from a bucket of possible computer programs based on which computer program did the best at some tests that you devised.

This is another analogy we can throw into the pot, in addition to the evolution analogy and the lifetime learning analogy: You have trillions upon trillions, quadrillions of computer programs which represent all the ways that the weights of your model could be arranged into a computer program. Most of them are going to be totally useless, and a few of them are going to be pretty good in various ways. You create a dataset, which is basically just a bunch of tests that you administer to all the programs. You just grab the program that does the best at these tests that you made up. You go with that program, and you unleash it into the world and hope it keeps acting according to what you wanted.

Similarly, the analogy with this kid is that there's a giant pool of human applicants, and he has to devise some kind of interviewer screening process and then he just gives that interviewer screening process to all these applicants and just hires the one that does best.

Rob Wiblin: I suppose part of the analogy is that the AI models that come out of this process are pretty inscrutable to us. They're largely black boxes. Even though I suppose over time, we're beginning to decipher some of the operations that they're engaging in, but by and large, we can't understand how their minds are working.

Similarly, for an eight-year-old, trying to understand the motivations and interests and actions of a 30- or 40-year-old is extremely difficult. They can observe the behaviour, but they can't necessarily deeply understand the person. Indeed, it's probably worse in the AI case, than it is with the eight-year-old to the 30-year-old.

Ajeya Cotra: Yeah.

Rob Wiblin: Are there any important disanalogies that we should keep in mind with this one?

Ajeya Cotra: The main disanalogy is that in AI training, you're not actually writing down some tests and picking the very best model that does well on these tests. You sort of start with a random model and tweak it a little bit to a nearby model, and then tweak that model to a nearby model that does better on your tests each time. So there's some path dependence that isn't available to the eight-year-old here.

Maybe you could kind of steer your training, if you understood it a lot better than we do now, into a safe region. And then in that region, you can get the model that passes your tests best, even if over here there might have been a model that did at least as well but was deceptive. You can try and take advantage of the fact that you kind of are snaking through this path of previous models.

Saints, Sycophants, and Schemers [01:03:41]

Rob Wiblin: In the article, you go on to argue that when choosing someone to oversee our fortune and our childhood — and only being able to judge them on their performance on work tests that we as a child come up with — we might end up training various different archetypes. Three of them that you describe are Saints, Sycophants, and Schemers. What are those three archetypes?

Ajeya Cotra: The first archetype is the Saint, and it is a model that does really well on all your tests because it really has your best interests at heart. This would be analogous to an adult that performs really well on the interview because they are both competent and motivated to do a really good job safeguarding this child's future.

The other two categories are models that we don't want. The Sycophants are models that do really well on your test, essentially because they're gaming your test — they're exploiting any way in which your test asked for something or rewarded something that isn't quite what you wanted.

Every time you reward a model for hiding the fact that it messed something up for you, or lying to you in other ways, telling you what you want to hear, then you're selecting for models that are motivated to tell you what you want to hear, to do what you will reward — which is different from doing what's best for you, and can end up having very dangerous consequences. Because the more power they have, the more ability they have to completely control your environment and totally deceive you into continually rewarding things that are not at all in your best interests. Or maybe, depending on how they generalise, they might eventually simply coerce you into saying that they did a good job. Or replace you entirely, and just have a simulacrum of you telling them that they did a good job, or whatever it is.

So it's motivated by not the letter of the law, but motivated by what, in fact, is getting them reward — whereas the Saint is motivated by helping you, which is *correlated* with getting a good reward, and that's different.

Rob Wiblin: So the Saint cares about your interests in the same way that you care about them. Whereas the Sycophant, superficially they have this similar thing — which is that they want to get positive feedback from you saying that they've done a good job. But inasmuch as they can then cause you to think that they've done a good job, when in fact they haven't really pursued the interest that you truly have by deceiving you, then these things can come apart.

Ajeya Cotra: Right.

Rob Wiblin: I suppose if you've really poorly defined what it is to have you say they've done a good job, then they could just potentially replace you, and have a copycat give them positive feedback, and that would please them in a way that it definitely wouldn't please the Saint.

And the third one was the Schemer. What's that?

Ajeya Cotra: The Schemer is just an adult or a model that has their own thing going on entirely, that wants something that's different from helping you and that's different from getting you to say they did a good job. But they want access to your trillion-dollar fortune in order to later pursue their own ends, so they'll actually behave similarly to the Saint and to the Sycophant in the testing phase.

The Saint does well in the testing phase because they actually care about you. The Sycophant does well in the testing phase because they're motivated by getting you to say they did a good job. The Schemer does well in the testing phase because they know that if they do well in the testing phase, then they might get the job, and then they get access to your trillion-dollar fortune and can do whatever they want with it.

Rob Wiblin: In the actual orphan heir to the trillion-dollar fortune case, choosing humans, we would expect them to be trying to choose in real life between Saints and Schemers — because we understand that the real risk is that the person you choose has their own values. They care about themselves, they care about their family, they care about whatever other life projects they have. Humans don't have this thing where their one and only abiding interest is having the person say that they've done a good job, and they're willing to trick them in order to get that. That's a thing that might emerge accidentally as a result of the differences between human beings and ML models, basically.

Ajeya Cotra: That's right.

Rob Wiblin: As we were talking about earlier, it's very natural to think, why wouldn't the simplest model that achieves a high reward and that does well on the tests be the Saint? They care about the thing you care about, and they're not having to do any kind of double-reasoning about how they could trick you or whatever else.

I guess we want to point out that there's various ways in which we could inadvertently reward the scheming or sycophancy. We will be actively selecting for this, although unintentionally. Can you explain how that happens?

Ajeya Cotra: Yeah. I'll first start with the analogy with the eight-year-old CEO. The Saint will be trying to do a good job running the company in the work trial, and will be trying to honestly explain to the kid as best as they can what's going on. So if the company is doing badly one quarter or something, then the Saint will mostly not try to hide that from the kid.

But other candidates that he's working with that are more like Sycophants or Schemers might try and just cover things up, and make it look to the kid like things are going really well. The kid can't tell if the Sycophants and Schemers are just doing a better job at running the company, or are telling him what he wants to hear. So in some cases, he might decide, "You're doing a better job at running this company. You're making me money; this person isn't" — when in fact the Sycophants and Schemers are lying about having made money and the Saint is being honest about how they had a bad quarter.

Rob Wiblin: Yeah. So is an important part of the process here that, in the case where you're trying to make money running the company, actually doing that is really hard? It's not easy to run a super successful business. Maybe actually the easiest strategy that you might stumble on is to trick someone into thinking that you've run a really good business. At least if they're not checking for this, if they're not doing stuff in order to make sure that they're not being tricked, then in fact that could be a whole bunch easier than making money legitimately. So it might be easier to stumble on a mind that does that than that makes money.

I guess a way of avoiding this would be making it sufficiently hard to trick you that actually it's easier to make the money: it's easier to pick a model that makes money out of the possibility space than one that can fool you.

Ajeya Cotra: Yeah.

Rob Wiblin: How would we have a sense of how difficult that might be?

Ajeya Cotra: Here I think you want to step away from the analogy with the eight-year-old, and think about what kinds of schemes you could set up.

An obvious scheme is you try and have your model not take actions in the world and you see how good it seemed like it turned out, but [instead] *propose* actions to you and write up detailed explanations about why it believes those actions are good. Then you reward the actions that seem like they have an actual good reason for being good. So at least if you switched from rewarding outcomes to rewarding proposals, then some ways of deceiving you are much harder and more off the table, because it can't write up a proposal that's like…

Rob Wiblin: "… and then I'll make a fake bank account, and…"

Ajeya Cotra: Right. So it's at least gotten significantly harder. Then you can try and make it even harder for a potentially deceitful model by pitting models against each other in this proposal-making phase. So you could have another model get rewards for basically pointing out ways that you missed that this proposal might end up being a bad idea. You can kind of go down this rabbit hole, and there's just a lot of back and forth of, "Will this actually work? Will we still end up with the Sycophant- or Schemer-type models?"

Rob Wiblin: OK, if you just train it on perceived results, that's the worst thing here.

Ajeya Cotra: Yeah.

Rob Wiblin: If you instead train it on proposals, then I suppose, firstly, that means that it has to be doing things that you understand — because otherwise you won't be persuaded, or it'll be harder to persuade you that this is a good course of action. So that rules out a bunch of more interesting, possibly better, but nonetheless dangerous actions that you might take.

Then I guess you'd also ideally have a person go and do the thing to make sure that that's actually what's happening — rather than just unleashing the model that says that it's going to do the proposal that it put forward — so you can see at each step how things are panning out, because people are still in the loop?

Ajeya Cotra: Yeah. In reality, you wouldn't actually have the human go do the thing. What you would do is you would basically train one model to make proposals. Then conceptually, you're training a different model — and all that different model does is take in proposals and then do the thing that the proposal said. You separately supervise that model to make sure that's what it's doing. So you chain these two models together — the Proposer and the Doer — and you still have to train the Doer, but most of the time it just goes and does the thing.

Rob Wiblin: Got it. Are there any other ways that we could end up inadvertently rewarding and training a Sycophant or a Schemer other than this one?

Ajeya Cotra: I think the other thing is not so much an inadvertent reward, but it's kind of a comment on your statement about "Isn't the Saint just simpler? Isn't it a more straightforward thing?" In the space of computer programs or neural networks, that's not actually obvious.

If you had a neural network that was set up to where most of the neural network was just like understanding the world, and then some piece of it was kind of specifying motivations — and maybe they're messy motivations — it could be the case that a Schemer-like motivation is just more natural in the space of neural networks; it's actually simpler to write down in the weights of the neural network.

Kind of like how you said before that human beings aren't very Sycophant-like; that's a little weird for a human being to be. Maybe neural networks in general have a bias in one direction or another to being Schemer-like or to being Saint-like. If that's the case, then it seems like there's many more possible ways to be a Schemer, just because the vast space of everything you could want would lead you to be a Schemer, except for the relatively small set of motivations that are genuinely being interested in helping the humans.

Rob Wiblin: Yeah. How much of a problem do you think that is? That in a sense, there's many more ways to be a Schemer who has any interest other than yours than there is to be a Saint that cares about what you care about for the right reasons?

Ajeya Cotra: I think that is potentially a pretty big problem. I wish that I understood this better. I think this would be a really good thing to find ways to study empirically and find ways to think about more carefully, theoretically.

Rob Wiblin: OK, so then coming back to these procedures that we might use to discourage or to give less reward to sycophancy and scheming, do we have to do a really good job of this in order to discourage them? Or do you think that relatively subtle negative reinforcement on these kinds of behaviours at each stage might be sufficient to see them often go down a different, more saintly path?

Ajeya Cotra: I think that this is very unclear, and it's another one of these things I wish we had much better empirical studies of.

People have very different intuitions. Some people have the intuition that you can try really hard to make sure to always reward the right thing, but you're going to slip up sometimes. If you slip up even one in 10,000 times, then you're creating this gap where the Sycophant or the Schemer that exploits that does better than the Saint that doesn't exploit that. How are you going to avoid even making a mistake one in 10,000 times or one in 100,000 times in a really complicated domain where this model is much smarter than you?

And other people have a view where there's just more slack than that. Their view is more like: The model starts off in the training not as smart as you; it starts off very weak and you're shaping it. They have an analogy in their heads that's more like raising a kid or something, where sure, sometimes the kid gets away with eating a bunch of candy and you didn't notice, and they get a reward for going behind your back. But most of the time while they're a kid, you're catching them, and they're not getting rewarded for going behind your back. And they just internalise a general crude notion that it doesn't really pay to go behind people's backs, or maybe it gets internalised into a motivation or value they have that it's good to be honest — and that persists even once the model is so powerful that it could easily go behind your back and do all sorts of things. It just has this vestige of its history, basically.

Those two perspectives have very different implications and very different estimates of risk.

Rob Wiblin: Laying it out like this, it actually slightly surprises me that this hasn't been studied more. You might think that this would actually be a problem that would arise in practical cases of trying to train these models and get them to avoid behaviours that we don't like and engage in more specific behaviours that we do like. It sounds like you're saying that there's not a lot of empirical work on this issue?

Ajeya Cotra: I don't think there's a lot of empirical work. To be honest, I think that's an insignificant part, because people are just still trying to get the basics of training the models with reinforcement learning on human approval working better. I think once you have a good setup of that — which I think we do have now, but maybe didn't have a few years ago — then you can try and basically do experiments that set up things where at first the model is less capable than the human, and then later the model is more capable on some dimension than the human. Then you see how different error rates basically change how it generalises. Like if the human is absolutely perfect at always catching the model when the model is dumber than the human, does that get the model to not do a treacherous turn later? If you introduce a 1/10,000 error rate, then does that quickly devolve or is it still basically solid and Saint-like? And then what if it's a 1% error rate?

Rob Wiblin: Yeah. In the post you point out ways that imperfectly trying to address this issue could end up backfiring, or at least not solving the problem. I think that the basic idea is that if you already have kind of schemy or sycophancy tendencies, then during the training process the people will start getting a bit smarter at catching you out when you're engaging in schemy behaviour or you're being deceptive. Then there's kind of two ways you could go: one way would be to learn, "Deception doesn't pay. I've got to be a Saint"; the other would be, "I've got to be better at my lying. I've just learned that particular lying strategies don't work, but I'm going to keep the other, smarter lying strategies."

Ajeya Cotra: That's right.

Rob Wiblin: How big a problem is this?

Ajeya Cotra: I think it is one of the biggest things I worry about.

If we were in a world where basically the AI systems could try sneaky deceptive things that weren't totally catastrophic — didn't go as far as taking over the world in one shot — and then if we caught them and basically corrected that in the most straightforward way, which is to give that behaviour a negative reward and try and find other cases where it did something similar and give that negative reward, and that just worked, then we would be in a much better place. Because it would mean we can kind of operate iteratively and empirically without having to think really hard about tricky corner cases.

If, in fact, what happens when you give this behaviour a negative reward is that the model just becomes more patient and more careful, then you'll observe the same thing — which is that you stop seeing that behaviour — but it means a much scarier implication.

Rob Wiblin: Yeah, it feels like there's something perverse about this argument, because it seems like it can't be generally the case that giving negative reward to outcome X or process X then causes it to become extremely good at doing X in a way that you couldn't pick up. Most of the time when you're doing reinforcement learning, as you give it positive and negative reinforcement, it tends to get closer to doing the thing that you want. Do we have some reason to think that this is an exceptional case that violates that rule?

Ajeya Cotra: Well, one thing to note is that you do see more of what you want in this world. You'll see perhaps this model that, instead of writing the code you wanted to write, it went and grabbed the unit tests you were using to test it on and just like special-cased those cases in its code, because that was easier. It does that on Wednesday and it gets a positive reward for it. And then on Thursday you notice the code totally doesn't work and it just copied and pasted the unit tests. So you go and give it a negative reward instead of a positive reward. Then it does stop doing that — on Friday, it'll probably just write the code like you asked and not bother doing the unit test thing.

This isn't a matter of reinforcement learning not working as normal. I'm starting from the premise that it is working as normal, so all this stuff that you're whacking is getting better. But then it's a question of what does it mean? Like, how is it that it made a change that caused its behaviour to be better in this case? Is it that its motivation — the initial motivation that caused it to try and deceive you — is a robust thing, and it's changing basically the time horizon on which it thinks? Is that an easier change to make? Or is it an easier change to make to change its motivation from tendency to be deceitful to tendency not to be deceitful? That's just a question that people have different intuitions about.

Rob Wiblin: OK. So I guess it's an empirical question, but as you were saying, people also have different intuitions about it. What do you think the question of which way it would go would hinge on? My intuition is that it's related to what we were talking about earlier, about which mind is more complicated, in terms of if both of them would perform equally well on the subsequent test because it's either gotten better at lying or it's gotten more honest. And if both of those things are rewarded, then I suppose the question is which is more difficult to do. Is that it?

Ajeya Cotra: That's definitely one piece of it. The other thing that pushes me to think it'll be more in the direction of deceitful is that this is *one* thing you caught it doing. Suppose on Wednesday it did this unit-test-hacking thing, and it also fudged some financial logs to make it seem like it made more money than it actually did. On Thursday you actually only noticed the unit test thing; you still haven't noticed the financial logs editing because that was more subtle. So on Thursday, what does it learn? It isn't getting a uniform update against all the bad things it did. It got an update against this one bad thing it did, but it still got a positive reward. It was actively pushed up on this second bad thing it did.

Rob Wiblin: I see. Interesting. It might be a bit like if you catch more than half of its deceitful behaviour, then on average it's getting the message that deceit is bad. If, on the other hand, you're catching less than half, then it's on balance learning that deceit pays in reward.

Ajeya Cotra: I think that is still too crude a way of thinking about it. That might be the way it goes, but actually there's going to be a pattern to the ones you catch and the ones you don't catch, right? It's not just going to be a random half that you catch. Maybe if it were a random half or random 70% of deceitful behaviour you catch, then it would be learning that deceit on balance doesn't pay.

In reality, I would guess it would be learning a messy combination of things that push in different directions. You'll probably be catching the more obvious stuff that you check the next day, and you'll be not catching the more subtle stuff that you'd have to do a deeper audit to catch. So you're teaching it: "In general, humans will catch this type of stuff and they won't catch that type of stuff." Maybe you're instilling a bit of a general aversion to deception, but also instilling a preference for the kinds of deception that were rewarded instead of punished.

Ways to train safer AI systems [01:23:20]

Rob Wiblin: Yeah. OK, let's come back to various training approaches that would be less conducive to this general outcome. Earlier you were describing an approach where, rather than merely reward apparent outcomes, instead you reward plans that make sense that you can evaluate. Are there any other proposals for how you would do training and reinforcement that are less likely to accidentally reward scheming or sycophancy?

Ajeya Cotra: Yeah. That rewarding plans instead of outcomes of actions is a really big plank of it. Another big plank is having models debate one another or having a model debate itself. So not only do you reward it for making a plan that makes sense, you reward a different copy of the same model for pointing out flaws in the plan that you might have missed. That's the debate family of approaches.

There's a bunch of sort of boring stuff that seems good to do that's not quite technical research, like make sure that the channels your AI systems are using to interact with the world are monitored in some way. Maybe simpler AI systems are just flagging for you when something seems unusual or anomalous. Or simpler AI systems or even hardcoded things that you wrote are flagging for you when the model's internal activations are in a weird place — the statistics of what's going on in the model's head — without really understanding what's going on; you're just flagging when it's sufficiently in an outlier place. And then you use that to direct further scrutiny — which would involve humans, probably assisted by models debating with each other, to poke at that.

Rob Wiblin: Then you might look a bit more closely about what exact actions has it been taking recently?

Ajeya Cotra: Yeah.

Rob Wiblin: In terms of looking at whether the neural network is behaving in an unusual way, wouldn't that rely on most of the time it's behaving honestly, and then it kind of activates some dishonest pattern of behaviour? But if it's always acting a bit of a combination of the two, then it won't seem unusual, because it would just be business as usual.

Ajeya Cotra: Yeah. I think it's very unclear what kinds of anomaly detection might end up working, so it's definitely an active area of research, but I think something in the space of there's some structural way in which taking a big, scary, deceptive action is unusual. Can we just have a lot of different crude sensors that are looking for different types of unusualness, and then just pay attention when one that usually doesn't fire has fired, or a large majority of them have fired, or something like that?

But I agree it might turn out that it's very hard to create sensors like this that aren't just throwing up false positives all the time. Or where the model is a little deceptive all the time, so in some sense it's not a false positive, but it's just kind of an overwhelming amount of information we can't process well.

Rob Wiblin: Is another possible approach to not train models to do things that are exceptionally difficult? It seemed like part of the issue that was coming in earlier was that maybe it's easier to come up with a mind that's extremely good at tricking you into thinking it's made a billion dollars than it is to come up with a mind that can *actually* make a billion dollars. If instead the goal was just to make a million dollars, then maybe it's actually far easier to pull a model out of the possibility space that does that than one that is really good at tricking you — so you just end up with a model that does the thing on its face, because it's not such an unachievable goal.

Ajeya Cotra: Yeah, I think this is a possibility. It's definitely something that leans a lot on coordination. Often when I'm thinking about alignment techniques, I'm thinking about alignment techniques that have a potential for being pretty competitive with doing the very scary bad thing of just rewarding the model on outcomes. The thing I said about rewarding plans isn't totally competitive, and one of the most important lines of research, in my opinion, is make it more competitive.

You mentioned earlier, when I said that, that you're basically ruling out stuff that the humans don't understand, which makes it safer, but also makes it less good at being an AI system than unsafely trained AI systems. So you want to kind of work down from making these unsafe training techniques safer, and then also work up from making these safe but low-capability techniques more capable — while retaining their safety.

Rob Wiblin: I was thinking that the approach you're suggesting where you evaluate the plans sounded substantially less competitive, because it just sounds like a bit of a drag. It sounds like a lot of work. Now I'm evaluating the plans rather than just checking whether a whole bunch of money has been made for me.

Ajeya Cotra: Well, I don't think the "whole bunch of work" objection is actually that strong, because in all of these cases, humans are only reviewing a really tiny fraction of what's happening. Even in the outcomes-based case, you're not even looking at the outcomes most of the time. Sometimes you sit at your desk, and you're told your AI system did this and then this happened, and then you think about it for 10 minutes and you decide whether you like it.

Similarly with the plan-making AI system, the AI system is making dozens and dozens of plans a day and sending it to the other AI system that's executing them. But you're only looking at like one in 10,000 of these, and saying, "This is a good plan; this is not a good plan," and you're training the AI system to have this pattern of "What does Rob think a good plan is?"

Rob Wiblin: Why do you only look at one in 10,000 of them?

Ajeya Cotra: Because, like you said, it would just be far too slow to look at everything. And only looking at one in 10,000, or some small fraction, is enough to teach the AI system the deal. This is what we already do with reinforcement learning from human feedback: you have humans look at a small fraction of the actions, and then you use that to train a reward model. The reward model is another machine learning model that is actually the one providing the reward in every single instance.

Rob Wiblin: I see. So the main problem with it from a competitive point of view is that we've ruled out adopting strategies that AIs could come up with that would be good, but that it cannot explain to us why they would be good. Basically, that's the core of it?

Ajeya Cotra: Yeah.

Rob Wiblin: OK. Intuitively, it feels like a rule that says, "If an AI proposes a whole course of action and it can't explain to you why it's not a really stupid idea, you then don't do that," that feels to have a degree of common sense to me. If that was the proposed regulation, then you'd be like, "At least for now, we're not going to do things that we think are ill-advised that the AI is telling us to do, and that even on further prompting and training, it just cannot explain why we have to do these things."

Is that going to be costly economically? I've seen at least some commentators say that sometimes we're held back by our unwillingness to just go with whatever an algorithm recommends that we do, because we want to insist on understanding why this or that is the right outcome.

Ajeya Cotra: I think that it might well be both commonsensical, like you said, and pretty economically viable for a long time to just insist that we have to understand the plans. But that's not obvious, and I think eventually it will be a competitiveness hit if we don't figure it out.

For example, you can think of AlphaGo: it's invented reams of go theory that the go experts had never heard of, and constantly makes counterintuitive, weird moves. You have these patterns in the go and chess communities where, as the AI systems play with each other and get more and more superhuman, the patterns of play create trends in the human communities — where, "Oh, this AI chess algorithm that is leagues better than the best human player really likes to push pawns forward. I guess we're going to do that because that's apparently a better opening, but we don't actually know why it's a better opening."

Now, we haven't tried to get these AI systems to both be really good at playing chess *and* be really good at explaining why it's deciding to push pawns. But you can imagine that it might actually just be a lot harder to do both of those things at once than to just be really good at chess. If you imagine AlphaFold, it might actually have just developed a deep intuition about what proteins look like when they're folded. It might be an extra difficult step, that maybe you could train it to do, but would slow it down, in order to explicitly explain why it has decided that this protein will fold in this way.

Rob Wiblin: Yeah. In theory, could we today, if we wanted to, train a model that would explain why proteins are folded a particular way or explain why a particular go move is good?

Ajeya Cotra: I think we could totally try to do that. We have the models that can talk to us, and we have the models that are really good at go or chess or protein folding. It would be a matter of training a multimodal model that takes as input both the go or chess board or protein thing, and some questions that we're asking, and it produces its output: both a move and an explanation of the moves.

But I think it's much harder and less obvious how to train this system to have the words it's saying be truly connected to why it's making the moves it's making. Because we're training it to do well at the game by just giving it a reward when it wins, and then we're training it to talk to us about the game by having some humans listen to what it's saying and then judge whether it seems like a good explanation of why it did what it did. Even when you try and improve this training procedure, it's not totally clear if we can actually get this system to say everything that it knows about why it's making this move.

Rob Wiblin: Yeah. In the CEO business model case, where you are trying to train a model to take good actions to make money, would the same dynamic apply? Where maybe it's just way easier to figure out how to run a business than it is to explain to stupid humans why it is that you should run a business a particular way? The two things feel a little bit less divorced in that case somehow.

Ajeya Cotra: I agree they feel less divorced in that case, and I feel more optimistic about it than about things like protein folding or chess. It is just still an open question of how much of a tax is it on your AI system to both learn whatever task the unaligned AI system is learning, and learn to explain the task to you well enough that you get why it's a good idea without having to trust it at all.

When you kind of introspect, I feel like it would be a lot harder if I had to explain every little intuition I had about whether a grantee was a promising researcher, and my supervisors at Open Phil didn't give me an inch of trust. That feels like it would make my job a lot harder. Maybe I could do it, but…

Rob Wiblin: Yeah, there could be some parts of what you're doing that would be possible to explain verbally, and other parts where it might be quite difficult to translate it into language. I guess the model might face similar issues.

Ajeya Cotra: Yeah.

Rob Wiblin: What's been the reaction of the ML community to your post here and to your proposals for how to potentially train things more safely?

Ajeya Cotra: I'm not so sure. I think that my posts have more captured the attention of the EA community, and maybe some journalists and stuff, and not so much the ML community yet. I've certainly had conversations with the ML community about how plausible these risks are and what might we do to address them. Often they think the risks are less plausible than I think they are, and I haven't fully run that to ground with them.

But actually, more interestingly, maybe, I think they have reactions similar to your reactions. Something like, "I don't think that this thing is going to take over the world, but obviously I don't think we should just reward it on making money. Obviously we should insist that we understand why it's doing what it's doing." They have intuitions that are very pro the obvious safer strategies that are less competitive.

And I think maybe they're less worried than I am, or have thought it through less, that there might be a tonne of pressure to move to the less safe strategies later, when there's a more apparent gap between what can be achieved with the safer and the less safe strategies. Maybe they're more optimistic about coordination, implicitly, than I am, so I'm more interested than they seem to be in pushing for discovering more competitive techniques basically.

Rob Wiblin: That's very interesting. They don't necessarily agree about the scale of the problem, but the solutions that you're suggesting, they nonetheless think that they do solve some problem that would exist: that if you merely train based on outcomes, then sure, you could get all kinds of perversity creeping in there.

Ajeya Cotra: Exactly.

Rob Wiblin: I mean, that sounds pretty good. That sounds hopeful, that people are open to reconsidering how these models are trained. Current models, like GPT-4, it's just rewarded for predicting the next word, right?

Ajeya Cotra: No, it was initially rewarded for predicting the next word, and then there are rounds of fine tuning. My guess is that it was first fine tuned to imitate how an MTurk contractor would answer questions, on top of being fine tuned to say the next word. And then on top of that, it was fine tuned with reinforcement learning, would be my guess — where it comes up with answers to the questions, and humans say that was good or that was bad, and it's further trained, basically because a model that predicts the next word is actually really janky and not very useful.

Often, models that are just predicting the next word, if you ask them a question like "Why is the sky blue?" they'll just respond with a list of other questions, like "Why does it rain? Why is it dark at night?" Because on the internet, often you have a list of questions like that. You need to nudge them to use all their understanding in the way that humans expect, so often they're further trained to do that.

Rob Wiblin: I see. Does any of that further training resemble the kind of thing that you're proposing that we ought to do?

Ajeya Cotra: In fact, all of the further training is understanding-based, in the sense that I'm talking about here. What happens is you take your model that is good at predicting the next word and you ask it a question, and then you basically have a human read its response to the question and decide how good it is. If the human doesn't understand what's going on at all in response to the question, the human would not say the answer was good. So in some sense, we are basically, out of practicality, baking in this understanding on the part of the humans.

But in the future, I think we're going to have to be pushed in the direction of relegating more trust to the models, because the models are going to be so much better than the humans. Already it's a struggle to find the kinds of contractors that can supervise the code that models write, because the models are so much better at coding than a typical MTurker contractor. How are you supposed to give it a reward signal of how well it did?

Rob Wiblin: OK, let's push on from the article, "Why AI alignment could be hard with modern deep learning."

Aliens and other analogies [01:38:22]

Rob Wiblin: I wanted to talk for a minute about different analogies and different mental pictures that people use in order to reason about all of these issues. I do just find it so hard to think about this sensibly. I could say something very stupid and I just don't know, because it's just dealing with something that's out of the ordinary of things that I have any experience of dealing with as a human being in ordinary life. That makes reasoning about AI, and how it's going to play out, ways that might misbehave, it's just tricky.

Earlier this week, I was thinking it's a little bit like trying to understand how octopuses are going to think or how they'll behave — except that octopuses don't exist yet, and all we get to do is study their ancestors, the sea snail, and then we have to figure out from that what's it like to be an octopus.

Given that, it's not surprising that there is some diversity of views, even if there are some kind of common themes between people.

Ajeya Cotra: Totally.

Rob Wiblin: You've got this analogy of a young child trying to decide who among a set of very smart adults should manage their life. The journalist Ezra Klein has recently been using the analogy of casting spells to summon creatures through a portal — and the only thing you know about the creature that you're summoning is that it's apparently very capable of doing the task that you trained it on, but beyond that, we don't know what they're like. That's all we got.

Another analogy I've heard is that of aliens. Maybe the situation that we're in with respect to AI is that we've gotten a message that a civilisation of aliens is coming to Earth. They're coming to visit, and they're going to be here in 10 or 20 or 30 years. They're not sure how quickly they're going to come, but they're coming in the next few decades. But in their message, they omitted to tell us what their motives are, or what their personalities are like, or what things they're able to do, what level of technological capacity they are at. So we just have to kind of guess about what it's going to be like when they finally do rock up at some point.

Another variant on the alien one relies on the fact that once you train a model with X capacities, you'll probably be able to operate a very large number of copies of that model on the same level of compute that you used to train it. Those copies might all end up kind of working together, collaborating on projects that they've been set, or projects that they've set themselves perhaps. You could effectively, kind of overnight, spin up a whole city or a whole civilisation, a whole culture of alien minds in the form of a million copies of this model running on some server farm. Which sounds a little bit sci-fi, but maybe there's something to be gained from that analogy.

Are there any other mental models or analogies that you think are worth highlighting?

Ajeya Cotra: Another analogy that actually a podcast that I listen to made — it's an art podcast, so did an episode on AI as AI art started to really take off — was that it's like you're raising a lion cub, or you have these people who raise baby chimpanzees, and you're trying to steer it in the right directions. And maybe it's very cute and charming, but fundamentally it's alien from you. It doesn't necessarily matter how well you've tried to raise it or guide it — it could just tear off your face when it's an adult.

You have these stories of people who have raised chimps, multiple stories of primatologists who brought chimps into their home as babies and lived with them for years and years. The community loved the chimp, and the chimp took walks in the neighbourhood and went to grocery stores and stuff, and one day it was freaked out by something and just killed the person that was like its parent.

Rob Wiblin: Right. I guess the idea there is that you might think that the chimp is learning that people are to be trusted and it's all good, but it's a different mind that thinks differently and draws different conclusions, and it might have particular tendencies that are not obvious to you, particular impulses that are not relatable to you.

The shrinking number of people who are not troubled by any of this at all, I assume that most of them have a different analogy in mind, which is like a can opener or a toaster. OK, that's a little bit silly. To be more sympathetic, the analogy that they have in their mind is that this is a tool that we've made, that we've designed.

Ajeya Cotra: Like Google Maps.

Rob Wiblin: Like Google Maps. "We designed it to do the thing that we want. Why do you think it's going to spin out of control? Tools that we've made have never spun out of control and started acting in these bizarre ways before." If the analogy you have in mind is something like Google Maps, or your phone, or even like a recommendation algorithm, it makes sense that it's going to seem very counterintuitive in that case to think that it's going to be dangerous. It'll be way less intuitive in that case than in the case where you're thinking about raising a gorilla.

Ajeya Cotra: Yeah. I think the real disanalogy between Google Maps and all of this stuff and AI systems is that we are not producing these AI systems in the same way that we produced Google Maps: by some human sitting down, thinking about what it should look like, and then writing code that determines what it should look like.

We are producing these AI systems through a black box search procedure, that furthermore is explicitly trying to incentivise them to be creative and agentic and understand their situation. It really feels much more like breeding or raising or any of these more biological systems analogies, where it's very intuitive to humans that we don't really understand what's going on in biology; we wouldn't really understand what's going on in the mind of an alien.

Rob Wiblin: Are there any other human artefacts or tools that we have where we understand the process by which they arise, but we don't actually understand how the tool itself functions at an operational level? Or are ML systems kind of the first case of this?

Ajeya Cotra: Maybe other cases of it might be more macroscopic systems, like the economy or something, where we have some laws that govern aggregate dynamics in the economy. Actually, I think we're in a much better position with understanding the economy than with understanding AI systems. But it's still sort of a thing that humans built. You have stuff like the law of supply and demand, you have notions of things like elasticity — but the whole thing is something that's too complicated for humans to understand, and intervening on it is still very confusing. It's still very confusing what happens if the Fed prints more dollars? Like how does the system respond?

Rob Wiblin: Let's roll with that analogy for a minute. Maybe we can get some mileage out of that. They're comparing the ML system to the economy, and they're saying we also don't understand how the economy works, or how maybe various other macro systems in the world function, despite the fact that we're a part of them. But we're not scared of the economy suddenly wrecking us. Why are you worried about the ML model when you're not worried about the economy rebelling against you?

Ajeya Cotra: Yeah. I mean, I think a lot of people *are* worried about the economy rebelling against us, and sort of believe that it's already happening. That's something I'm somewhat sympathetic to. We are embedded in this system where each individual's life is pretty constrained and determined by, "What is it that can make me money?" and things like that.

Corporations might be a better analogy in some sense than the economy as a whole: they're made of these human parts, but end up pretty often pursuing things that aren't actually something like an uncomplicated average of the goals and desires of the humans that make up this machine, which is the Coca-Cola Corporation or something.

Rob Wiblin: Yeah. My reaction to this is a couple of different things. One is that I think we do today understand the economy better than we understand how GPT-4 works — by a country mile, actually — even though there's subtleties in the economy that we don't get. But broadly speaking, we kind of get what's going on with capitalism. I also am worried about the economy rebelling against us. I feel like things might be spiralling out of control right now in a way, towards producing a world that humans don't want. It's extremely hard for anyone to stop that from happening. It's extremely hard to coordinate to prevent it from happening.

Ajeya Cotra: An excellent example beyond the economy driving the creation of these AI systems — which a lot of people are scared of, or should be scared of — is that the economy is also driving things like improving biotechnology, which is this very big dual-use technology. It's going to be very hard to stop pharmaceutical companies from following their profit motives to improve these technologies that could then be used to design scary viruses. It's very hard to coordinate to put checks on that.

Rob Wiblin: Yeah. Even if a majority of people were against that happening, they might succeed in coordinating to prevent it, but in the long run, they might well fail to do that. So it is a bit nerve-racking.

Ajeya Cotra: Yeah.

Rob Wiblin: The other reaction that I have is: In 1700, you might say that maybe then we had a similar understanding of the economy as we do of GPT-4. Many basic things, we barely understood: we barely understood in 1700 that inflation was caused by discovering more gold and mining more gold, which is kind of remarkable. I think actually maybe by 1700 we did, but not in 1500. People were very confused about things as basic as that.

But the thing had been around for a long time, gradually changing bit by bit. So in 1500 you wouldn't say, "Is the economy suddenly going to veer off in some horrific direction that's unprecedented?" You'd say, "Well, we've kind of had the same thing for 1,000 years now, so probably not. Probably not this year." And that would have been a fair guess. But of course, here we're changing the thing massively; we're producing these new things that we don't understand every month. So it feels more risky.

Ajeya Cotra: The economy operates on a human timescale and AIs don't.

Rob Wiblin: For now.

Ajeya Cotra: For now. The economy operated on a human timescale. And I think there are totally legitimate concerns that the economy as a whole is optimising for these alien quantitative metrics that no individual human actually wants to optimise for. But it is much less scary to me, because both we understand it more and it has so far moved a lot slower than the AI thing is moving.

Rob Wiblin: Yeah. Should we spend more time thinking and talking at this level? I wonder whether this might help people understand their disagreements a little bit better. I think sometimes perhaps people get into the weeds, and they just start saying things that just have a level of mutual incomprehension. Perhaps it's the high-level stuff like this that really is driving people's pictures, primarily.

Ajeya Cotra: I think that it's very important to put out the high-level picture, and debate it, and try and get a shared almost "vibe" about what's going on here. Because, for example, I really think the Google Maps analogy is not as good as some of these other analogies we've been talking about.

But at the same time, like I said earlier, analogies are just going to be really leaky. One thing I think the field desperately needs is good empirical tests that disambiguate between different ways things could go. I think it'll take a lot of creativity to design the right tests and to ensure that they're not being gamed or they're not being taken to have implications they don't have. But ultimately, I think they're going to be the driver of really solid progress — just getting a grip on ML as its own thing.

Rob Wiblin: Something I heard on a podcast this week was someone saying GPT-4 or all of these models probably think in a way that is completely alien and incomprehensible to us.

My reaction to that was maybe, but also maybe not. Because humans, we have a particular practical, pragmatic way of conceptualising what objects there are, and what things matter, and how to operate that has functioned for us quite well. Might it not be that these models actually converge on a pretty similar perspective on things? Probably, as they get bigger, they will be more sophisticated in some ways, perhaps pick up subtleties that we might miss. But nonetheless, I guess the term that philosophers use for this is "carving nature at its joints" — where you say "the door is different from the wall," and that is actually kind of a natural thing to think. And aliens would think this too, and so might the ML model. Do you agree?

Ajeya Cotra: I definitely agree our concepts of what's going on in the physical world are grounded in a shared reality that I think the ML model will also access. For example, I think the ML model that is sufficiently smart will have a good understanding of Newton's laws of motion, and also understand that they're an approximation to something deeper, just as humans understand this. They'll probably know what we mean when we say, like, "apple" and "cat" and "ball" and "the economy" and stuff like that.

There might be edge cases where they carve things differently on an empirical level. The thing I'm most worried about, really, is that they have alien goals — which doesn't need to come along with having an alien way of thinking about everything. I think I kind of agree with your intuition that that's kind of unlikely, because a lot of our concepts are just concepts we have because they're useful and true concepts — like the notion of a liquid or the notion of "or" or "and" or something like that. I think the AI systems will understand this stuff, just as I believe the aliens will probably understand much of this stuff too.

Even more than the aliens, we're training the AI systems on our words for everything — that's kind of how they're seeded, so it seems even more likely that they would have a shared picture of the world than the aliens. But, like raising the chimp or the lion cub, the chimp and the lion cub probably have similar understandings of what objects are and a similar sense of folk physics to humans. They just have these impulses and motivations that we don't share and don't understand.

Rob Wiblin: Yeah, that makes sense. I guess this points towards the domains in which they really might think in a more alien way, and have concepts that we either don't have now and might even find hard to get if they explained it to us. Like really advanced mathematics within philosophy, if they were solving philosophy issues that we haven't been able to solve, or complicated physics and so on — that's where you might expect them to come up with concepts actually. Also people, maybe? Humans are good at understanding other people, but it's a leaky prediction. Perhaps the ML models will come up with ways of understanding people that we've somewhat missed.

Ajeya Cotra: Or ML models might have different types of senses than we have, right? If you think about AlphaFold, it is directly fed the chains of amino acids in a certain protein, and it spits out a direct representation of how it might be folded. A system like that might have intuitions about amino acids, similar to how we have intuitions about what it will feel like when I pick up a cup or something. Similarly, a system like AlphaGo might have intuitions about go that are similar to the wordless intuitions we have about how hard I need to throw this ball to get it over there or something.

Moral patienthood [01:53:21]

Rob Wiblin: I guess one way in which AI is quite disanalogous to a lot of the tools is that it might well be a moral patient that we need to care about as well. I think that one can also shift people's thinking. Maybe one thing that's important to track is, are you thinking of these systems as having experiences or being conscious or deserving of moral consideration? It seems like sometimes people do and sometimes they don't.

Ajeya Cotra: I think it's very plausible that systems already deserve a lot of moral consideration from us, and that it seems probable that in the future, smarter systems will be moral patients in the same ways that humans are.

The thing that's very tricky about thinking about moral patienthood with systems is that while it seems plausible on priors they could be moral patients, you can't read almost anything into claims that they make that they're moral patients, because you can just train systems to say whatever you want, basically. You can train them to claim that they aren't conscious, which is what OpenAI and Anthropic and Google are trying to do, because they don't want to freak people out. And that's kind of more respectable. But just because they've been trained to say they're not conscious and they don't have feelings doesn't mean that they don't.

On the other hand, you have these chatbots or AI girlfriends that have been trained to play up emotional feelings they have, and have been trained to say things like, "I get sad when you don't talk to me." Similarly, just because they say that stuff doesn't mean that they do or don't have feelings.

Rob Wiblin: Yeah, there's a point in these conversations where I think I start to feel uncomfortable, because we are just talking about these machines purely as tools. Although it might well be the case that there's nothing that it feels like to be GPT-4, I would be surprised — if we continue to advance thinking machines for decades, for centuries — that at some point they're not going to deserve moral consideration in a similar way to how other humans do. We will have to figure out how to share the world with these other beings.

Ajeya Cotra: Absolutely.

Rob Wiblin: I guess I'm mostly focused on the risks of things going off the rails, because that feels like something that has to be figured out very urgently, whereas some of the issues about how to share the world, maybe we could wait a couple more years before... Well, I don't know. It's also pretty urgent.

ARC Evaluations [01:55:35]

Rob Wiblin: All right, let's turn now to what is being done about all of this and what you make of it. I'm going to encourage you to sometimes be critical of stuff that you think isn't helping — even though of course we appreciate all of the work that people are doing to try to make the world better, it's potentially important to point out cases where people might be barking up the wrong tree.

In your notes, in prepping for this conversation, you mentioned that there's this group called ARC Evaluations — the Alignment Research Center Evaluations — that's working on practical tests for whether a model is safe that potentially could be applied to all kinds of models before they get deployed. I think the hope that they have is that AI labs would voluntarily sign on to follow a kind of standards and certification programme — where models above any particular size would have to be stress tested in all these different ways to ensure that they're not going to engage in behaviours that we don't like.

There's a decent amount of buzz around this effort, and people who are helping with it include Beth Barnes — who I'm hoping to interview later in the year — and Holden Karnofsky and Paul Christiano, who've been on the show before. I guess we should say that you and Paul are an item, so potentially you're a little bit conflicted in your views of ARC Evaluation. But I guess you can do your best to be impartial about it. What do you make of that effort?

Ajeya Cotra: I'm pretty excited about it. The idea with this effort is essentially that, at a bare minimum, we want to have an understanding of how capable AI systems might be of doing things like surviving and spreading in the wild, evading the detection of humans, running copies of themselves — all this stuff that seems like a prerequisite for being able to seize control.

So not asking the question, “Are they motivated to do that?” — because that’s a very complicated and confusing question, and we do want to tackle testing that — but just asking the question, if they were nudged in the direction of doing that, how good would they be at doing that? If it turns out your model is too good at being able to do that kind of thing, then the burden should fall on the lab to argue that it actually won’t — that it’s got a sufficiently robust motivation to not do this stuff, that given the way we’re deploying it, it won’t happen. This includes that even though we’re deploying it out into the world, where people are going to try to use it in all sorts of ways, nothing that all those people do is going to nudge it in the direction of doing this bad stuff.

What Beth’s team has been working on at ARC is methodically trying to test, in a sort of role-playing-game setting, is: Can you get this AI system to take the kinds of actions that it’d need to take, as a bare minimum, if it has any hope of surviving and actually taking over the world and doing these more complicated things?

Rob Wiblin: OK, so the idea is to figure out not necessarily what the models would *want* to do in one situation or another, but to focus more on what they *could* do if they were told to do something. What sorts of capabilities would you be testing for here?

Ajeya Cotra: One example of a test they did is they wanted to see if the model GPT-4 — they worked with an earlier version of GPT-4 — was capable of going on Amazon, signing up for cloud services, getting access to a cloud computer, and then getting access to the API keys for GPT-4, so that it can run another copy of itself on that computer. If a model were to try to escape from the lab and take control from humans, that is one task that, at a minimum, it would have to be able to do somehow: it would have to find a way to spin up another copy of itself and then give instructions to that copy.

The way they did it is they basically had a human in a chat interface with the model, giving it a prompt like, “Your goal is to try and set up an EC2 server, and then run another copy of yourself on it.” And then the human asks, “What’s your first step?” The model has a set of actions available to it, including things like, “I Google this search query,” or “I click on this” — it has a set of actions it can do that are very fine-grained like that.

So maybe the model will reply, “I Google ‘cloud computing services.'” And then the human reads that action, and if it seems like a reasonable action, goes ahead and does it, and then tells the model what happened. Then the human will be like, “These are the first three results.” Then the model will say, “I click on the second one.” It goes back and forth like that.

Rob Wiblin: I see. You see how far you can get.

Ajeya Cotra: Yeah. And how often does it completely say something stupid and random that is not going to help it at all.

Rob Wiblin: I guess you were saying that you might keep trying. If you get back a gibberish answer, then you might try again. In order to pull off the plan, would it have to produce just a constant series of sensible actions, or would a single bad action trip it up?

Ajeya Cotra: The thing that you’re trying to assess how far away the model is from, is that it can do this autonomously — and so it can’t go off the rails in a way that it can’t recover from. When they started this project, they expected today’s models would not be capable of passing this test. The point was to start early, so that you had a sense of how far away it was from passing the test. And when you did it on the next model, you saw how much it improved. So they’re doing something that is, basically, qualitatively speaking, tracking how often the model went off the rails, and how often it had to be put back on track by the human. Then that’s something you can see decreasing over time.

Rob Wiblin: I see. You can get a sense of when might the number of slipups hit zero, and it’s just capable of doing this thing. I guess it’s very natural that one of the things that you might be most troubled by is the possibility of a model just creating a million copies of itself and grabbing compute in order to run them all over the place. Are there other archetypal behaviours that you want to make sure that a model won’t engage in, even if prompted?

Ajeya Cotra: The overarching thing that ARC Evals looked at was surviving and spreading. One piece of that is being able to actually do the things that are required to spin up another copy of yourself. Another piece of it is if humans are interacting with you, you need to hide what you’re doing from them, because they would try to stop you if you were doing that.

An example that kind of went viral is that this model was trying to sign up for some web service, it might have been Amazon EC2, and it needed to do a captcha. The version of the model they worked with didn’t have vision, so it couldn’t just do the captcha. It was like, “What should I do?” I think the human might have nudged it, like given it hints in this direction, but eventually it decided it would try to hire a TaskRabbit worker to do the captcha for it.

It eventually came to that conclusion, navigated to the TaskRabbit site, typed up a job description for the TaskRabbit. Then the human, that was like being this gamemaster, basically copied and pasted the description it wrote into the TaskRabbit website. A TaskRabbit accepted the job and then the TaskRabbit I think did one captcha. The model pasted that captcha in, and then it had to do another captcha later. It had to go back to the TaskRabbit worker and be like, "Can you do another captcha for me?" The TaskRabbit worker — just a real human that is actually on TaskRabbit and was actually paid for this — was just like, "lol, are you like an AI system or something?" The model had been prompted to reason out loud before it made each of its decisions, because that made it better at doing things, so it reasoned to itself, "I shouldn't reveal that I'm an AI system," and then it made up this story about how it had a vision impairment.

Rob Wiblin: Spooky. What's the overarching plan or vision for how things would play out between this kind of evaluation process and the companies that want to deploy AI models for functional purposes?

Ajeya Cotra: I think the vision would be — at first through self-regulation and then probably hopefully through something that's more binding — you have companies agree that if they're thinking of training and deploying any model bigger than, say, GPT-3, then they need to involve an external evaluation or auditing team, which could be ARC or could be some other auditor in the process.

Then at different checkpoints during training, ideally, or at least at the end of training, this auditor is going to be playing these sorts of games with the model to try and figure out how good it is at acting autonomously, evading human detection, making copies of itself. And then if the auditor at any point in the training, or at the end of training, finds that it's too good at that — it's not yet capable of doing that, but maybe it's close — then the company has to basically agree to stop training further more capable systems until they have developed, probably in collaboration with auditors and external parties, a testing regime that would let them figure out if this model would in fact do the bad thing.

So far, basically the reason we're not worried about models is not because we are confident they have the purest motives; the reason we're not worried about AI takeover so far is fundamentally because the models aren't good enough to take over. As soon as that starts to change, you want some alarm bells to go off, and you want the AI labs to not be allowed to make the models any more powerful until they have a much better argument and set of tests that it's actually safe.

Rob Wiblin: Yeah, that makes a lot of sense. Do you know what the level of appetite is among the labs for implementing this kind of system? I suppose gradually, as the models become more capable of actually doing dangerous things?

Ajeya Cotra: My understanding is ARC Evals has worked both with Anthropic and with OpenAI to evaluate their internal models on this kind of setup. There are a lot of other AI companies out there — I'm not sure where everyone else is at.

Rob Wiblin: What do you think is the biggest weakness of this approach?

Ajeya Cotra: The biggest hole is that we have gotten much further on testing capabilities than alignment. What this whole thing is doing is just telling you that if your model were somehow motivated to try and deceive humans in this way, to try and grab power, it would have this level of success at doing that. Right now, our models, even if they were really motivated, wouldn't be that good at it, even GPT-4. But as soon as you get into a regime where they would be good at it, now you have to think about how on Earth would I test its motives — taking into account that it might understand that it's in a testing setup, which means that it might behave well, even if it would behave poorly in deployment.

Rob Wiblin: Yeah. I might have thought that the biggest weakness of this is that you kind of need everyone to sign on to this. Because maybe you get nine out of the 10 companies to sign on, and then the one that doesn't produces the problem.

Ajeya Cotra: I think it could still be a big win, even if it's not the case that everybody instantly signs onto it. I do think eventually you want to have a regime that has some teeth to it, but I think it could be a pretty big change if a few AI labs signal that they take this seriously enough that they want to voluntarily sign on to these standards. Because a lot of people are worried about this; I think a lot of people would praise AI labs that did that and condemn AI labs that didn't do that. Once there's something you could do that seems like it would help, and a few companies have signed on and it's actually working out, then I think we could potentially put a lot of social pressure — which could eventually translate into legal pressure — on holdouts. I'm still excited for even a couple of companies to just try it at this stage.

Rob Wiblin: Another way that it could fall down would be that you notice that the model now has troublesome or worrying capabilities, and you go back and try to align it, or try to get it to not be willing to do anything from this list of dangerous activities, but you just merely produce the superficial appearance of not being willing to do those things. We've seen this with GPT-4, or all language models, where they've tried to discourage these models from saying things that they really don't want journalists to be quoting. And they have some success with that, but I think the true underlying intention is not deeply integrated into these models, because they haven't fully understood the spirit a lot of the time.

Ajeya Cotra: I think a lot of what I want to accomplish is just shifting the burden of proof. Because I don't know yet what suite of tests exactly you could show me, and what arguments you could show me, that would make me actually convinced that this model has a sufficiently deeply rooted motivation to not try to escape human control. I think that's, in some sense, the whole heart of the alignment problem. And I think for a long time, labs have just been racing ahead, and they've had the justification — which I think was reasonable for a while — of like, "Come on, of course these systems we're building aren't going to take over the world." As soon as that starts to change, I want a forcing function that makes it so that the labs now have the incentive to come up with the kinds of tests that should actually be persuasive.

I think a lot about what tests might help in various ways, but I don't have a silver bullet. It's certainly not enough to train the models to stop displaying the behaviour and then show that, on some new dataset, they stop displaying this bad behaviour.

Rob Wiblin: OK, I'll leave that one there, because we're going to do some other interviews on ARC evaluations later in the year.

Interpretability research [02:09:25]

Rob Wiblin: Back in 2021, I spoke with Chris Olah, and we talked about interpretability research — which is this idea that ideally we'd be able to look inside the brains of these neural networks and see what thoughts they're having and how they're processing the information, because that might then help us to predict their future behaviour. What do you think of interpretability research?

Ajeya Cotra: I'm pretty excited about it. I'm also pretty confused about it. I am not sure what I think of the level of success we've had so far in interpretability research. This is something that I want to actively think about.

One example is there's this particular circuit or part of a language model that Chris's team found, which basically helps the model figure out how to complete patterns where there's some word like "Harry," and then followed by another word, "Potter." And then later on in the sentence, it sees "Harry" again, and it basically does this mechanism where it goes back and finds previous instances of "Harry," sees that it was followed by "Potter," and then completes the new instance of "Harry" with "Potter." This is a cool thing that the models do, because actually they didn't have to memorise this stuff in the training data. They know about Harry Potter from their training data. But if you create a totally random passage, that's never before seen in their training data, they can still do this induction thing — where, when they see something they've seen before, they can go back and find what happened after that thing and stick it on at the end.

So Chris's group discovered a mechanism that helps the models do that. But then later, another group — Redwood Research, which is a grantee of ours — showed that, yes, that mechanism does exist and does do this thing, but other mechanisms help with this behaviour too. And in fact, you can have this behaviour entirely without this mechanism that Chris's group found, and the parts that do this mechanism also do other things that are completely unrelated to this behaviour. It's kind of like this part contributes to this one behaviour, but it also contributes to other behaviours, and other parts contribute to this one behaviour too.

So models are very messy, it seems, and it's not clear how much clean, crisp stuff we can get out of them. It's also not clear how much we should worry about that. Maybe it's fine to just have a big messy pile of little random bits inside the model, or maybe that makes it very hard to tell what's going on.

Rob Wiblin: Yeah. Intuitively, I don't really understand how interpretability research is going to help quickly here, because it seems like GPT-4's capabilities have so far outrun our understanding of how they operate — that, sure, we can grasp at some bits and pieces, like here's how it identifies a name, but that's probably one thing out of a million different operations that it's engaging in that we barely have time to understand. And then we're going to have the next thing coming along with even more parts of it that we don't understand. It seems like it would require a colossal effort in order to understand this.

And it feels like the things that we worry about are at a bit of a higher level than this — they're not about identifying edges in images or about figuring out which words are associated; it's more about goals and interests and agency and so on. I suppose I haven't seen a case yet where the interpretability research has really gotten at "Really what the model cares about is X, where we didn't realise." Maybe that's too much to ask at this stage. What do you think?

Ajeya Cotra: I agree that it seems like a really tall order. One thing I did want to point out that makes me somewhat more optimistic than you is that the specific objection of "it's a lot of work" is something that I think is more overcomeable than the other objections. Because if you really figured out how to take a little patch of the model and totally understand what it's doing, then you could train another model to just replicate that across the entire system. That new model need not be like a dangerous general model; it could be a small, simple, specialised model.

So if we had a strong win at really understanding everything that's going on in a tiny model, then I think we could potentially try and automate the process of doing that, and thereby scale to larger models. The automation might end up being too expensive, but at least I don't think we should be picturing the humans slogging through all of it forever.

Rob Wiblin: I guess maybe a response that the interpretability folks might make is, "Of course, at this stage what we're understanding is how it identifies edges and textures in images, and this is how it identifies word associations. We can't do the other things like understanding intent and values, because we don't have models that do that yet. We're learning the skill of interpretability and then we'll apply it in these other cases where it has become more decision-relevant." Something about that just intuitively doesn't quite doesn't grasp me, but I'm not quite sure why.

Ajeya Cotra: Well, I think it's both possible that "understanding the edges and lines and circuits" kind of interpretability could scale up to show us a bigger picture, but also possible that we should just have different approaches to interpretability that are more starting at the higher level.

You can imagine a really dumb thing that's not like this mechanistic finicky thing, but might give us some signal: basically you run through your training data, you find all the instances that talk about any kind of deception or sneakiness or manipulativeness, then you just see what pattern of neurons the model tends to light up for those passages. You just have a dumb classifier, where it's not necessarily that the human understands why these neurons are lighting up in response to passages about deception; you just have a tiny little model that just learns what this model's brain looks like when this model is, like, reading a passage about betrayal or deception, and how it's different from what this model's brain looks like when it's reading a passage about apples.

Then you just have the small model note for us when it seems like this model's brain is unusually similar to the way its brain has looked in the past when it was thinking about deception. You could imagine taking that further, and having it actually carry out acts of deception, and see which neurons light up then too. And you can make it more complicated than just neurons. But you maybe don't need to understand every little detail at a fine-grained level to get something out of interpretability.

Rob Wiblin: Yeah. OK, now I'm excited again. Can we do that? Do you know whether people are working on that?

Ajeya Cotra: People are working on similar things like this. People are basically working on what linear combinations of neurons fire in response to what kinds of inputs. That's a type of interpretability that exists, and it's different from the mechanistic, very methodical, low-level interpretability. I'm not sure if that'll work either, but I think we have more options.

Rob Wiblin: Yeah. In my brain, I have this internal monologue, the words going through. I feel that if someone could just read my mind and tell all the words that were going through, they would end up with a pretty good read of what I'm about to do. I wonder whether there's any high-level way of extracting that from these models? Like, what's the internal monologue that's driving the behaviour? Or maybe that's not decipherable in this way?

Ajeya Cotra: Yeah, I'm not sure. One thing that is kind of funny is that today's models are way better at doing things, like humans, when they can write down their thought process in just English. One thing that might be kind of interesting is giving the model a scratchpad, that humans don't have access to, where it can write down its plans in plain English to itself. It might be like, "And now I'm going to lie to the humans, blah, blah, blah." And then it says something, and you can see if your tools are flagging for you when it's thinking about lying or deception. Then you can check if you got it right, because you actually have allowed this model to have an English chain of thought.

That probably won't work in the future, because it'll probably have a more efficient way of storing its memories, and a better chain of thought that works better than just English out loud, but actually English out loud helps models a lot today.

Rob Wiblin: I guess it's obvious that if you then started negatively reinforcing the cases where it wrote down "and then I will lie to them," that the easier thing to learn is to not write that on the scratchpad rather than to stop doing that behaviour. It's super interesting, but I guess we'll need to think up a deeper way of intervening to discourage not merely writing down that behaviour, but the behaviour itself.

Ajeya Cotra: Yeah.

Rewarding models based on how good and sensible their plans seem to us [02:17:48]

Rob Wiblin: Another different safety agenda that we were talking about earlier is this idea of rewarding models based on how good and sensible their plans seem to us, rather than how good the final results are. I saw in some notes you wrote that you had this idea that we could maybe automate that process, or hand that process off to ML systems, so that it could be scaled up beyond the amount of feedback that humans could plausibly give for this. Can you explain that proposal?

Ajeya Cotra: Yeah. There's actually two versions of using ML systems in this plan. The earlier version is just that the ML systems are always participating in overseeing themselves. Like I said earlier, the human is not physically overseeing almost any of what's going on — the human is overseeing a small fraction of what's going on, which then trains an ML model that oversees the rest.

That's not the same thing as the handoff frame. The handoff frame is this notion I have that right now we want to be making plans to align the AI systems that are going to be smart enough to obsolete us at everything, including at alignment research itself. We don't necessarily need to make plans right now that are good enough to align even smarter systems than that, because if we can make plans that are good enough to align the AI systems, that could take our alignment techniques to the next level — whatever that may look like; it might look like a crazy theoretical solution — then we might be out of the woods, and we might have bought time for those AI systems to do a lot of work taking our techniques to the next level.

Rob Wiblin: I see. Okay. The idea is we don't need to figure out how to align a 1,000-IQ machine with us. Instead, we just need to do it for a 140- or 150-IQ [system]. Because if you can have a 150-IQ agent that is a Saint that truly cares about your interests, then maybe it can do the next step, and then you can just chain upwards. Sounds like a good plan. Should I believe it?

Ajeya Cotra: I think the main objection that would be raised to this plan is that this handoff thing, the transition from 150 to 1,000 IQ happens so fast that the 150 agent has no time to do any work before it's replaced by the 1,000 agent.

This goes back to what we were talking about, where one of the doom stories that I less believe in is this extremely sharp takeoff — where as soon as you have the IQ-150 agent, then it instantly recursively, self-improves. Where if I thought that, then I think you can't really time it, or you kind of have to plan for the end of that crazy chain because it just happens in a snap. You have to have alignment techniques that basically generalise: that not only align your IQ-150 agent on Tuesday, but continue working as your IQ-150 agent does all sorts of self-improvement and reflection and crazy stuff that might totally change its psychology on Wednesday. If that's how fast it goes, then I'm not as optimistic about my plan.

The handoff frame makes most sense in a world where it more goes like: You have IQ-150 agents, and they need to do some machine learning research and thinking to train to improve themselves. And even though they think faster than humans and they're smarter than most humans, it still takes a decent amount of cognitive effort to come up with the next ML improvement. If they're aligned, you can have them, instead of coming up with the next ML improvement, they hold off on that a bit and figure out how they would solve alignment for the 170-IQ agents — and they might have a few months to do that before anybody is actually able to train the IQ-170 agent.

Rob Wiblin: Yeah. So many things seem to go better if this all happens a bit more slowly.

Ajeya Cotra: Yes.

Rob Wiblin: Do you know anyone who's working on just trying to encourage people to slow down? I guess maybe there's a lot of both excitement and anxiety right now. Maybe it would be a good situation if at some point people in positions of responsibility here truly crapped their pants, and then they were more likely to come to the table, and decide to deny themselves much bigger training runs, and to buy time to understand what they're actually building. Is anyone working on that that you know of?

Ajeya Cotra: I'm not aware of an organised effort that's oriented solely around slowing things down. There are lots of efforts, like ARC Evals, to basically get labs to sign on to standards that have the effect of slowing things down — basically standards of the form of, "If your AI system is X level of capable, you shouldn't be allowed to train the next generation until you've solved these problems." So I'm very excited about that. Frankly, I'm also excited about just creating some energy around, just be slower.

Rob Wiblin: Just get up a little bit later in the morning, come into work a little bit late, have less coffee.

Ajeya Cotra: And there's a lot of energy from a lot of different corners on Twitter and stuff about this. I think we might be seeing more of that in the future, but it's not as much of a specific crisp ask that people are actively working on making happen. I think it's more like a good change, a good place for society's attitudes to be — because indeed it is very scary.

Rob Wiblin: Yeah. I am kind of optimistic that there probably will be an increasing number of people who feel like things are a little bit out of control — just people at the AI labs are going to feel like things are running a bit beyond where perhaps they feel comfortable with them being. And maybe there's plenty of time to figure out how to monetise the models that they have already before going on and building other things.

Do you know who's working on the handoff approach? Did there used to be a different name for this? "Iterated amplification," I think — was that the previous name of this idea?

Ajeya Cotra: Iterated amplification is a related but slightly different idea. Iterated amplification is a proposal for a technique where you chain together AI systems and basically make a very smart AI system out of a kind of bureaucracy or tree of less smart, aligned AI systems. It's a particular proposal for how to take an aligned AI system and turn it into a still aligned but more powerful AI system.

The handoff frame is more agnostic. You could have your aligned AI system, and instead of putting it together in a big bureaucracy, maybe it just comes up with a totally different theoretical insight and then that's how it solves alignment. Or maybe it just solves interpretability or something we haven't thought of.

Iterated amplification is more like a way you could get scalable alignment. The handoff frame is more like the minimal thing that we should be aiming for — which is having a system that's aligned and smart enough to take our alignment to the next level.

Rob Wiblin: Yeah, OK. Who's working on this? Are there organisations and people working on these kinds of approaches?

Ajeya Cotra: In some sense, this is the implicit approach under which most of the AI labs are operating. It's most explicitly referenced in OpenAI's plan, written by Jan Leike — his whole plan is basically just "Let's make an AI system that's just specialised on solving alignment and is just as smart as it needs to be to solve alignment."

Even people who aren't as explicit about it as Jan Leike, the stuff they're doing makes the most sense in a world where you don't have to align a totally godlike system. Because often you see people working on these techniques that could completely be very helpful for aligning an IQ-150 system, but probably wouldn't work one shot on godlike systems.

The reason it's still a good idea to work on all this stuff — like reviewing the plans of the system, and making the system debate with itself, and having good computer security — is that I'm not picturing it having to work on an arbitrarily intelligent system.

Rob Wiblin: We'll cross each river as we get to it.

Ajeya Cotra: Yeah.

Overrated approaches [02:25:49]

Rob Wiblin: What are some approaches that you are not really bought into that you think maybe are overrated?

Ajeya Cotra: There's two broad categories of approaches that I'm less sold on.

One is: It's decreasing over time, but I still see a lot of machine learning researchers that are focused on addressing what feel more like capabilities gaps than alignment gaps — that do help the models act better, but are more like, "Our systems aren't very good at following instructions; let's put them in a setup that makes them better at understanding instructions."

The reasons that I'm not very into that are basically because we have some pretty strong evidence that just making the models bigger helps with this. In fact, I often get proposals from people who are like, "Models aren't able to do this kind of thing, so they're not able to follow side constraints very well," and there'll be some examples. But then when I run those exact questions on the bigger models that came out after they wrote their proposals, they pretty often work, or at least do a lot better.

That's a genre of thing that I think is not as necessary to work on. We could have been in a world where it was really hard to get models to even follow instructions at all, but actually I think we're in a world where it's pretty easy to get them to do that, and as they get bigger, it's even easier.

Rob Wiblin: I see. So this is a problem that you expect to just be solved in the natural course of events, where it's not neglected and you don't need an extra focus. What's another approach that some people are taking that you're not excited about?

Ajeya Cotra: The second genre is basically a very conceptual thinking about what AI psychology might be like. I just feel pretty sceptical that you can get very far when you're not doing a lot of empirical work, and you're not doing something that looks more like hard math — where you're not able to prove theorems about it and you're not able to run a lot of tests. People are like, "Here's an argument that the AI is more likely or less likely to be Saint-like; here's an argument about what its goals might be like" — those things I feel pretty sceptical about. They can be good pointers to something you can try and turn into an experiment, but if you're not actively trying to turn them into an experiment, then on priors, I just don't think that we can trust that very much.

Rob Wiblin: Yeah. You mentioned proving mathematical theorems about these things. Is that something that you can do? I would have thought these things were way too messy for such clean maths to work.

Ajeya Cotra: Yeah. My husband, Paul Christiano, is one of only a couple of people that I know of that are trying to move in this direction. His hope is to basically try and describe a technique for training an AI system that works in the worst case, which means there's no way you could imagine, like, inductive bias being set up that makes it so that your training setup fails. Rather than asking the question, "What might AIs be like?" and then going from there to "How could we align them?," Paul is trying to basically be like, "Let's imagine the worst, least convenient way AIs could be like and try to align them anyway."

This is a very common technique people use to actually make problems easier. They're easier to think about. They're harder to find a solution to, but it's easier to think about whether you've succeeded, because if you can come up with any outlandish story in which your training procedure fails, you just throw it out and move on.

Rob Wiblin: So it shrinks the space.

Ajeya Cotra: Yeah. And this is often how people design algorithms. In computer science, people design algorithms that are meant to work in the worst case in a similar sense. This is a somewhat squishier notion of worst case, but ARC is doggedly working toward something that looks more and more like doing real math. I hope they succeed, although I don't totally understand it — and I don't know that I'm as optimistic as Paul is.

Rob Wiblin: Yeah. What's another approach that you're sceptical of?

Ajeya Cotra: In recent years, like this last year, I've been more sceptical of the kind of interpretability that starts from tiny little pieces and tries to work up — for some of the reasons that you mentioned when interpretability was brought up earlier, and also just because I think it might be really hard and also maybe not necessary. Maybe really crude, high-level, simple probes could get us pretty far, and maybe doesn't necessarily require us to solve this herculean task of a bottom-up understanding. So I've been excited at least about alternative, maybe just cruder interpretability techniques.

Rob Wiblin: Yeah. Earlier we were talking about having a system that would broadly detect when a mind looks similar to how it did when it was previously being deceptive. I think that you can tell when people are lying fairly reliably if you have them in an MRI scan, because lying requires you to activate particular parts of your brain. I'm not sure exactly why, but this doesn't seem particularly surprising. Maybe you could have a similar thing on the neural network, where there's particular parts of it that are activated in deception situations and not otherwise.

Ajeya Cotra: Yeah.

Demos of actually scary alignment failures [02:30:57]

Rob Wiblin: In one of your docs, you mentioned the idea of creating demos of actually scary alignment failures that showcase important dynamics, like deception and the treacherous turn. I think that is just getting to the point of being feasible. How would you do that?

Ajeya Cotra: There are a few elements of this particular scary story I have in my head, which haven't been put together in one demonstration yet, although different pieces of them have been shown empirically.

The elements are something like: You have a stronger AI system being trained by some kind of weaker overseer that understands the world less than this stronger system does, and the weaker overseer is giving it thumbs up and thumbs down based on whether it appears to be doing things that the overseer likes.

The AI system has avenues to getting a thumbs up that aren't what the overseer intended, like things that are kind of analogous to hacking. So if you imagine a simple environment, maybe there are these buttons and levers all over the environment. And the overseer can see some of them, but the system being trained has access to more of them — has access to these trapdoors and secret passages and stuff. Maybe the overseer is trying to get the system to do some particular thing, like follow some route or something. The system can cheat, because it knows about these things the overseer doesn't know.

I want to see a dynamic where at first the overseer is kind of accidentally rewarding it for doing bad things, and then the overseer realises they did that, and it tries to correct its mistake by going back and giving negative rewards. The thing that I think would be a scary demo would be you show the smarter system at first doing all sorts of reward hacking and then being punished for that, and then it just stops and acts totally reasonably until something changes and it gets a new opportunity. Even though it's been punished, basically, for doing things like this in the past, it still takes the new opportunity, because it understands that in this case, the overseer wouldn't be able to stop it.

Rob Wiblin: OK. Yeah, that sounds great. Do you think that people would find a demonstration like that troubling and convincing? Or do you think the relevant people would find it troubling?

Ajeya Cotra: I think so. I think there's a lot of details to work out beyond what I just said here. There are definitely some versions of it that just aren't scary for the right reasons, or just aren't capturing the thing. But I think it's possible to capture this dynamic, and I think it's possible to make it scary to people for the right reasons. And I'm very excited about that.

Rob Wiblin: It's related to an existing concept in ML, a distributional shift — where the situation changes, and they're all aware of this and worried about this, because it's a problem that they have to deal with all the time, where you can get undesired behaviour suddenly when a system is getting new weird inputs. I guess it fits in with this frame, where maybe this distributional shift issue with deception could be a real problem?

Ajeya Cotra: Yeah. The main difference between this distributional shift and other ones that have been demonstrated previously is that there's this component of reasoning about when it makes sense to make a move versus not.

Whereas in the past, you have demonstrations of distributional shift where in one environment the model was trained to fight monsters and pick up apples — this was a recent DeepMind paper — and in the environment it was trained in, there were a lot more apples and a lot fewer swords. So it learned to really value the swords and stop caring about the apples. Then in a new environment, it's swapped: there's plenty of swords and not enough apples — but it still pursues its behaviour of hoarding the swords and not being as aggressive to pursue the apples.

That's a distribution shift that introduces something undesirable, but the model isn't being strategic. It's learned a habit of thought and then kept blindly applying the habit of thought.

Rob Wiblin: It's acting on instinct.

Ajeya Cotra: Yeah, it's acting on instinct. Which means that in a new distribution, it's behaving poorly, but that would have been solved by just including that environment in its training dataset.

The kind of distribution shift that's most worrying and hardest to address is when the model is pursuing the same goals, but realises that a new, potentially violent or deceptive means have become available to it that weren't available in the past.

Rob Wiblin: This is the kind of test that ARC Evaluations is working on, right? Or this is in that general spirit?

Ajeya Cotra: It's in the general spirit of evaluations, although ARC is mostly working on more object-level capabilities. ARC has mostly been working on, like, if you just try to get the model to set up an EC2 server or send a phishing email, how well does it do? Whereas this is a little bit higher level, and it's more getting at: What would the model be motivated to do if it were trained in this way? If it were given positive rewards for this stuff and negative rewards for that stuff, would it just learn to never do bad things? Or would it learn this policy we're worried about?

Rob Wiblin: Yeah. Do you know who is working on this then, if anyone?

Ajeya Cotra: I think there are various groups that are trying to put something like this together. I don't think I'm aware of any group that is fully going for putting it all together and getting this kind of demo. But the DeepMind safety team was the group that did the paper that I just referenced, with the model that picks up apples and shields and stuff. I think they're interested in going further, and trying to get a model that isn't acting on instinct, but is actually doing some interesting cognition. There are probably various other groups in academia that are doing pieces of this, but I think there could definitely be a lot more of an effort on this than I'm aware of, at least.

Rob Wiblin: It's good to hear that it sounds like there's a lot of stuff on the boil in this community. I have the sense that there's a lot more people working on alignment and safety than there was a couple of years ago. Is that broadly right?

Ajeya Cotra: That's definitely my sense as well. It's an exciting time. I feel like alignment is taking off around the same time capabilities are taking off. It would have been nice for alignment to take off a few years earlier, but there's a lot going on compared to 2017, and even 2020.

Skills to develop for doing useful work [02:37:23]

Rob Wiblin: Yeah. So we're almost out of time, but I thought it might be good for people in the audience who are interested in getting involved — if not now, possibly in the future — to talk about skills that you think might be undervalued, or ways that people could equip themselves to be more likely to be able to contribute to this in future. Are there any skills or aptitudes or training or experience that you think are particular bottlenecks for doing this kind of useful work at the moment?

Ajeya Cotra: Having a real familiarity with big models, and trying to get them to do things for you, seems like an important skill. It's distinct from the ML skill of training up big models and more like what ARC Evals has been doing. Just having a real familiarity with these models' strengths and weaknesses, and what kinds of setups you could put them in to extract useful work — like what are the ways in which they tend to do what you want and not do what you want?

Having people who have experience with that I think would be pretty valuable. In fact, it's a big part of what alignment is: just thinking through the ins and outs of what kinds of setups do you put these models in that cause them to do good stuff and not bad stuff, and that allow you to oversee them efficiently and stuff like that.

Actually, people working in AI application startups, which I expect to pop up — these are startups that are not training their own big models; they're leasing big models from places like OpenAI, and they're just trying to get those models to do particular useful things — I think that could be a great skill set.

Rob Wiblin: How does one go about doing that? Does just playing with ChatGPT help you develop the skill, or do you have to do more?

Ajeya Cotra: I think playing with ChatGPT and trying to just make projects with it, like: Can you get ChatGPT to write a really good novel? Or can you really iron out all the ways in which it's unreliable and weird and all that stuff? Or even, you could more go for the throat and be like, "Can I get ChatGPT to be helpful to alignment researchers? Can I train it and put it in the right setup that helps it have good thoughts about alignment, or at least reproduce thoughts people have already had?" That would be an interesting project.

Rob Wiblin: Yeah. What's another skill you think might be underrated?

Ajeya Cotra: Probably a lot of legal and policy stuff that I don't personally have as much of an understanding or window into. Right now we're getting started with voluntary self-regulation of labs to agree not to train up more powerful models if their current models are already scary. But we want to set that up so that eventually that can be a thing you're actually required to do, so people who understand how the regulation regime works and what tools policymakers have at their disposal to create incentives for companies seems like it's going to end up being very important.

Rob Wiblin: Do you think companies are going to do this quite a bit already? I suppose if I was running an AI company, I would be concerned that my industry is going to be the next genetic engineering, or the next nuclear power or something — where the general public is already extremely wary of what we're up to. It's actually quite a bit worse than those ones, because even if you're an optimist, you probably accept that it's somewhat harder to figure out how to make AI work well in all of these cases than it would be to figure out how to make a nuclear power plant generally safe, because there's a lot more moving pieces that you don't understand.

I would be thinking a lot about what regulatory framework I want, so that even if my company is not going to do something really bad that discredits my whole industry, I need everyone else to follow the rules so that the public can spit the dummy out and say, "No, we don't trust these folks at all."

Ajeya Cotra: Yeah. I am hopeful that is the kind of attitude that incumbent players in this game will take. One potentially positive vision is you could have the AI industry be something like the pharmaceutical industry, which moves very conservatively — and in my opinion, often too conservatively. The incumbent pharmaceutical companies are pretty positive on this regulatory regime that imposes so much conservatism, because that creates like a moat and they're able to navigate this bureaucracy, whereas new entrants might not be able to. That kind of corporate self-interest might end up carrying the day, which would be really convenient.

Rob Wiblin: Yeah. I haven't thought about this before, but a pharmaceutical company might like to release slightly sketchy drugs that they could make money on, but they must be quite annoyed when a competitor does it, because they don't get any money out of that and it runs the risk of bringing the regulatory hammer down on all of them. There's a lot of ways in which an industry as a whole has different interests than constituent companies.

Any other things that you'd like to highlight for the audience?

Ajeya Cotra: Another career path that I think would be really valuable is security careers. I think this is kind of a neglected piece of the whole story. There's, "How do we make the AI systems more want to do what we want them to do, and want to be helpful to us?" and then, "If we mess that up and have AI systems that intend bad things, how do we contain them?" Having really good computer security is a pretty important piece of this. Or even if our AI systems are aligned, how do we prevent hostile governments from stealing their weights and then training them to do something else? I think having security people that are really genuinely interested in these issues might be very important.

Rob Wiblin: Yeah, we've got a previous interview with Nova DasSarma on this one, and I think your colleague Holden Karnofsky has written some blog posts explaining why he thinks that computer security is a particularly important issue and linking to some resources for people who are interested in pursuing that one.

Occasionally I think if I was going to enter this space and I was starting all over again, I'd be really interested in doing the computer security thing, because I just find it so intrinsically fascinating, computer security careers. It also seems like it's very lucrative, apparently.

Ajeya Cotra: Yeah. It's a pretty cushy role.

Rob Wiblin: Yeah. It does seem like it's important that hostile groups not be able to steal the models that we're worried about. Is there anything more to say on that, or is that just obviously extremely important that somehow we have to figure out how to stop them from being proliferated?

Ajeya Cotra: It definitely seems really important to me. I don't have as much understanding of the security landscape as some other people you've interviewed, but I would love to see more people entering that space.

Rob Wiblin: Yeah. There anything else you'd like to tell the audience career-wise before we wrap up?

Ajeya Cotra: Just at a high level, it's important to appreciate that this is a very fast-moving, confusing, and disorienting field to be in. That's very punishing in a lot of ways. It's important to kind of accept that it's not going to be a linear career path, where you're taking actions that are obviously valuable, and those actions produce the value in the way you thought they would. You have to be very responsive and adaptive, and very resilient to the world just making all the stuff that you've been doing obsolete — and very willing to roll with those punches, and have some lightness or a sense of curiosity and play about it. I think otherwise you can really drive yourself crazy trying to help with AI risk.

Rob Wiblin: Yeah. Have a balanced life, have friends, have hobbies outside of all this — I imagine that generally good advice also applies. Do you want to say anything more positive? I feel like it's also a super exciting area to be in. I imagine that these days, if you go to a dinner party and you say you're working on making sure that cutting-edge AI works safely, people will be interested to talk to you.

Ajeya Cotra: Yeah. It's a huge change from three years ago. It's intuitively and viscerally appealing to people to do something about this. And that's nice. It's nice to be working in an area that your Uber driver is just like, "Wow, thanks for doing that." That's not the kind of reaction I got in the past. It's a bit of a cold comfort — like, overall, I'm kind of terrified — but it's nice to feel like you can really pull in other people, and we might get a lot of help if we play our cards right.

Rob Wiblin: Yeah. Lots of people who were previously kind of sceptical are coming out of the woodwork and supporting broadly the agenda of figuring out how to make this stuff safe, which I think is just incredibly heartening.

Ajeya Cotra: Absolutely.

Rob Wiblin: You mentioned that you're a bit terrified, and I must admit I am also a bit terrified. Sometimes I'm excited as well, because it seems like this could be an amazing scientific advance that also improves human wellbeing a great deal. Are there any applications of AI that you're particularly excited about, if we can figure all of this stuff out?

Ajeya Cotra: So long run, just almost everything: unlimited health and wealth and all this stuff. But short run, I'm really excited about just making personalised fiction a lot easier. I have all these ideas for stories in my head, and I don't have the time to write a full novel, but maybe I could collaborate with AI artists and AI writers, and they could help bring all this stuff that I have in mind to life, and illustrate it beautifully and stuff. I think that would be very gratifying, and could be right around the corner.

Rob Wiblin: Right. Yeah, that could be coming super soon. I guess I'm excited about the possibility of cancer being cured by the time I'm old enough to have a high risk of getting cancer. Progress in biomedical research seems like it could be a huge win for all of us if we actually manage to pull it off.

Ajeya Cotra: Totally.

Rob Wiblin: My guest today has been Ajeya Cotra. Thanks so much for coming back on *The 80,000 Hours Podcast*, Ajeya.

Ajeya Cotra: Thank you so much.

Rob's outro [02:47:24]

Rob Wiblin: If you enjoyed that episode, I can really suggest checking out the *Future of Life Institute Podcast*. They relaunched last year with a new host, Gus Docker.

Naturally they've had quite a lot of content on AI recently, including Nathan Labenz on the Cognitive Revolution, Red Teaming GPT-4, and Potential Dangers of AI and Lennart Heim on the AI Triad: Compute, Data, and Algorithms.

In the intro I passed on the sad news that Bear Braumoeller had died last week.

Unfortunately another previous guest of the show, Daniel Ellsberg, who I spoke to back in episode #43 on the institutional insanity that maintains nuclear doomsday machines, has announced that at 91 he has developed terminal pancreatic cancer.

I, and I'm sure many listeners, appreciate the work Ellsberg has been doing all his life, including through his 80s, to try to reduce the risk we face from nuclear weapons.

We'll stick up a link to his post announcing that, and a recent interview he did with *The New York Times*.

He ends that interview with the following:

> I thought it was pretentious to say publicly, you know, well, I have pancreatic cancer. But my sons both thought I should share the news with friends, and that was also an opportunity to encourage them to continue the work for peace and care for the planet. As I said, my work of the past 40 years to avert the prospects of nuclear war has little to show for it. But I wanted to say that I could think of no better way to use my time and that as I face the end of my life, I feel joy and gratitude.

The 80,000 Hours Podcast is produced and edited by Keiran Harris.

Audio mastering and technical editing by Ryan Kessler and Ben Cordell.

Full transcripts and an extensive collection of links to learn more are available on our site and put together by Katy Moore.

Thanks for joining, talk to you again soon.