Lex Fridman Podcast #452 - Dario Amodei: Anthropic CEO on Claude, AGI & the Future of AI & Humanity

Published - November 11, 2024

Transcribed by - thepodtranscripts.com

Lex Fridman
The following is a conversation with Dario Amodei, CEO of Anthropic, the company that created Claude, that is currently and often at the top of most LLM benchmark leaderboards. On top of that, Dario and the Anthropic team have been outspoken advocates for taking the topic of AI safety very seriously. And they have continued to publish a lot of fascinating AI on this and other topics. I'm also joined afterwards by two other brilliant people from Anthropic. First Amanda Askell, who is a researcher working on alignment and fine-tuning of Claude, including the design of Claude's character and personality. A few folks told me she has probably talked with Claude more than any human at Anthropic. So she was definitely a fascinating person to talk to about prompt engineering and practical advice on how to get the best out of Claude.

After that, Chris Olah stopped by for a chat. He's one of the pioneers of the field of mechanistic interpretability, which is an exciting set of efforts that aims to reverse engineering neural networks, to figure out what's going on inside, inferring behaviors from neural activation patterns inside the network. This is a very promising approach for keeping future super-intelligent AI systems safe. For example, by detecting from the activations when the model is trying to deceive the human it is talking to.

This is the Lex Fridman podcast. To support it, please check out our sponsors in the description. And now, dear friends, here's Dario Amodei.

Let's start with a big idea of scaling laws and the scaling hypothesis. What is it? What is its history, and where do we stand today?

Dario Amodei
So I can only describe it as it relates to my own experience, but I've been in the AI field for about 10 years and it was something I noticed very early on. So I first joined the AI world when I was working at Baidu with Andrew Ng in late 2014, which is almost exactly 10 years ago now. And the first thing we worked on was speech recognition systems. And in those days I think deep learning was a new thing. It had made lots of progress, but everyone was always saying, "We don't have the algorithms we need to succeed. We are only matching a tiny fraction. There's so much we need to discover algorithmically. We haven't found a picture of how to match the human brain." And in some ways it was fortunate, you can have almost beginner's luck. I was like a newcomer to the field. And I looked at the neural net that we were using for speech, the recurrent neural networks, and I said, "I don't know, what if you make them bigger and give them more layers? And what if you scale up the data along with this?" I just saw these as independent dials that you could turn. And I noticed that the models started to do better and better as you gave them more data, as you made the models larger, as you trained them for longer. And I didn't measure things precisely in those days, but along with colleagues, we very much got the informal sense that the more data and the more compute and the more training you put into these models, the better they

perform. And so initially my thinking was, "Hey, maybe that is just true for speech recognition systems. Maybe that's just one particular quirk, one particular area." I think it wasn't until 2017 when I first saw the results from GPT-1 that it clicked for me that language is probably the area in which we can do this. We can get trillions of words of language data, we can train on them. And the models we were trained in those days were tiny. You could train them on one to eight GPUs, whereas now we train jobs on tens of thousands, soon going to hundreds of thousands of GPUs. And so when I saw those two things together, and there were a few people like Ilya Sudskever who you've interviewed, who had somewhat similar views. He might've been the first one, although I think a few people came to similar views around the same time, right? There was Rich Sutton's bitter lesson, Gwern wrote about the scaling hypothesis. But I think somewhere between 2014 and 2017 was when it really clicked for me, when I really got conviction that, "Hey, we're going to be able to these incredibly wide cognitive tasks if we just scale up the models." And at every stage of scaling, there are always arguments. And when I first heard them honestly, I thought, "Probably I'm the one who's wrong and all these experts in the field are right. They know the situation better than I do, right?" There's the Chomsky argument about, "You can get syntactics but you can't get semantics." There was this idea, "Oh, you can make a sentence make sense, but you can't make a paragraph make sense." The latest one we have today is, "We're going to run out of data, or the data isn't high quality enough or models can't reason." And each time, every time, we manage to either find a way around or scaling just is the way around. Sometimes it's one, sometimes it's the other. And so I'm now at this point, I still think it's always quite uncertain. We have nothing but inductive inference to tell us that the next two years are going to be like the last 10 years. But I've seen the movie enough times, I've seen the story happen for enough times to really believe that probably the scaling is going to continue, and that there's some magic to it that we haven't really explained on a theoretical basis yet.

Lex Fridman
And of course the scaling here is bigger networks, bigger data, bigger compute?

Dario Amodei
Yes.

Lex Fridman
All of those?

Dario Amodei
In particular, linear scaling up of bigger networks, bigger training times and more and more data. So all of these things, almost like a chemical reaction, you have three ingredients in the chemical reaction and you need to linearly scale up the three ingredients. If you scale up one, not the others, you run out of the other reagents and the reaction stops. But if you scale up everything in series, then the reaction can proceed.

Lex Fridman
And of course now that you have this kind of empirical science/art, you can apply it to other more nuanced things like scaling laws applied to interpretability or scaling laws applied to post-training. Or just seeing how does this thing scale. But the big scaling law, I guess the underlying scaling hypothesis has to do with big networks, big data leads to intelligence?

Dario Amodei
Yeah, we've documented scaling laws in lots of domains other than language. So initially the paper we did that first showed it, was in early 2020, where we first showed it for language. There was then some work late in 2020 where we showed the same thing for other modalities like images, video, text to image, image to text, math. They all had the same pattern. And you're right, now there are other stages like post-training or there are new types of reasoning models. And in all of those cases that we've measured, we see similar types of scaling laws.

Lex Fridman
A bit of a philosophical question, but what's your intuition about why bigger is better in terms of network size and data size? Why does it lead to more intelligent models?

Dario Amodei
So in my previous career as a biophysicist… So I did a physics undergrad and then biophysics in grad school. So I think back to what I know as a physicist, which is actually much less than what some of my colleagues at Anthropic have in terms of expertise in physics. There's this concept called the one over F noise and one over X distributions, where often, just like if you add up a bunch of natural processes, you get a Gaussian, if you add up a bunch of differently-distributed natural processes… If you take a probe and hook it up to a resistor, the distribution of the thermal noise in the resistor goes as one over the frequency. It's some kind of natural convergent distribution. And I think what it amounts to, is that if you look at a lot of things that are produced by some natural process that has a lot of different scales, not a Gaussian, which is kind of narrowly distributed, but if I look at large and small fluctuations that lead to electrical noise, they have this decaying one over X distribution. And so now I think of patterns in the physical world or in language. If I think about the patterns in language, there are some really simple patterns, some words are much more common than others, like the. Then there's basic noun-verb structure. Then there's the fact that nouns and verbs have to agree, they have to coordinate. And there's the higher-level sentence structure. Then there's the thematic structure of paragraphs. And so the fact that there's this regressing structure, you can imagine that as you make the networks larger, first they capture the really simple correlations, the really simple patterns, and there's this long tail of other patterns. And if that long tail of other patterns is really smooth like it is with the one over F noise in physical processes like resistors, then you can imagine as you make the network larger, it's kind of capturing more and more of that distribution. And so that smoothness gets reflected in how well the models are at

predicting and how well they perform. Language is an evolved process. We've developed language, we have common words and less common words. We have common expressions and less common expressions. We have ideas, cliches, that are expressed frequently, and we have novel ideas. And that process has developed, has evolved with humans over millions of years. And so the guess, and this is pure speculation, would be that there's some kind of long tail distribution of the distribution of these ideas.

Lex Fridman
So there's the long tail, but also there's the height of the hierarchy of concepts that you're building up. So the bigger the network, presumably you have a higher capacity to-

Dario Amodei
Exactly. If you have a small network, you only get the common stuff. If I take a tiny neural network, it's very good at understanding that a sentence has to have verb, adjective, noun, but it's terrible at deciding what those verb adjective and noun should be and whether they should make sense. If I make it just a little bigger, it gets good at that, then suddenly it's good at the sentences, but it's not good at the paragraphs. And so these rarer and more complex patterns get picked up as I add more capacity to the network.

Lex Fridman
Well, the natural question then is what's the ceiling of this?

Dario Amodei
Yeah.

Lex Fridman
How complicated and complex is the real world? How much is the stuff is there to learn?

Dario Amodei
I don't think any of us knows the answer to that question. My strong instinct would be that there's no ceiling below the level of humans. We humans are able to understand these various patterns. And so that makes me think that if we continue to scale up these models to kind of develop new methods for training them and scaling them up, that will at least get to the level that we've gotten to with humans. There's then a question of how much more is it possible to understand than humans do? How much is it possible to be smarter and more perceptive than humans? I would guess the answer has got to be domain-dependent. If I look at an area like biology, and I wrote this essay, Machines of Loving Grace, it seems to me that humans are struggling to understand the complexity of biology. If you go to Stanford or to Harvard or to Berkeley, you have whole departments of folks trying to study the immune system or metabolic pathways, and each person understands only a tiny bit, a part of it, specializes. And they're struggling to combine their knowledge with that of other humans. And so I have an instinct that there's a lot of room at the top for AIs to get smarter. If I think

of something like materials in the physical world, or addressing conflicts between humans or something like that, I mean it may be there's only some of these problems are not intractable, but much harder. And it may be that there's only so well you can do at some of these things. Just like with speech recognition, there's only so clear I can hear your speech. So I think in some areas there may be ceilings that are very close to what humans have done. In other areas, those ceilings may be very far away. I think we'll only find out when we build these systems. It's very hard to know in advance. We can speculate, but we can't be sure.

Lex Fridman
And in some domains, the ceiling might have to do with human bureaucracies and things like this, as you write about.

Dario Amodei
Yes.

Lex Fridman
So humans fundamentally has to be part of the loop. That's the cause of the ceiling, not maybe the limits of the intelligence.

Dario Amodei
Yeah, I think in many cases, in theory, technology could change very fast. For example, all the things that we might invent with respect to biology, but remember, there's a clinical trial system that we have to go through to actually administer these things to humans. I think that's a mixture of things that are unnecessary in bureaucratic and things that kind of protect the integrity of society. And the whole challenge is that it's hard to tell what's going on. It's hard to tell which is which. I think in terms of drug development, my view is that we're too slow and we're too conservative. But certainly if you get these things wrong, it's possible to risk people's lives by being too reckless. And so at least some of these human institutions are in fact protecting people. So it's all about finding the balance. I strongly suspect that balance is kind of more on the side of wishing to make things happen faster, but there is a balance.

Lex Fridman
If we do hit a limit, if we do hit a slowdown in the scaling laws, what do you think would be the reason? Is it compute-limited, data-limited? Is it something else? Idea limited?

Dario Amodei
So a few things, now we're talking about hitting the limit before we get to the level of humans and the skill of humans. So I think one that's popular today, and I think could be a limit that we run into, like most of the limits, I would bet against it, but it's definitely possible, is we simply run out of data. There's only so much data on the internet, and there's

issues with the quality of the data. You can get hundreds of trillions of words on the internet, but a lot of it is repetitive or it's search engine optimization drivel, or maybe in the future it'll even be text generated by AIs itself. And so I think there are limits to what can be produced in this way. That said, we, and I would guess other companies, are working on ways to make data synthetic, where you can use the model to generate more data of the type that you have already, or even generate data from scratch. If you think about what was done with DeepMind's AlphaGo Zero, they managed to get a bot all the way from no ability to play Go whatsoever to above human level, just by playing against itself. There was no example data from humans required in the AlphaGo Zero version of it. The other direction of course, is these reasoning models that do chain of thought and stop to think and reflect on their own thinking. In a way that's another kind of synthetic data coupled with reinforcement learning. So my guess is with one of those methods, we'll get around the data limitation or there may be other sources of data that are available. We could just observe that, even if there's no problem with data, as we start to scale models up, they just stopped getting better. It seemed to be a reliable observation that they've gotten better, that could just stop at some point for a reason we don't understand. The answer could be that we need to invent some new architecture. There have been problems in the past with say, numerical stability of models where it looked like things were leveling off, but actually when we found the right unblocker, they didn't end up doing so. So perhaps there's some new optimization method or some new technique we need to unblock things. I've seen no evidence of that so far, but if things were to slow down, that perhaps could be one reason.

Lex Fridman
What about the limits of compute, meaning the expensive nature of building bigger and bigger data centers?

Dario Amodei
So right now, I think most of the frontier model companies, I would guess, are operating in roughly 1 billion scale, plus or minus a factor of three. Those are the models that exist now or are being trained now. I think next year we're going to go to a few billion, and then 2026, we may go to above 10 billion. And probably by 2027, their ambitions to build hundred billion dollar clusters. And I think all of that actually will happen. There's a lot of determination to build the compute, to do it within this country, and I would guess that it actually does happen. Now, if we get to a hundred billion, that's still not enough compute, that's still not enough scale, then either we need even more scale, or we need to develop some way of doing it more efficiently of shifting the curve. I think between all of these, one of the reasons I'm bullish about powerful AI happening so fast, is just that if you extrapolate the next few points on the curve, we're very quickly getting towards human level ability. Some of the new models that we developed, some reasoning models that have come from other companies, they're starting to get to what I would call the PhD or professional level. If you look at their coding ability, the latest model we released, Sonnet 3.5, the new or updated version, it gets something like 50% on SWE-bench. And SWE-bench is an example of a

bunch of professional real-world software engineering tasks. At the beginning of the year, I think the state of the art was 3 or 4%. So in 10 months we've gone from 3% to 50% on this task. And I think in another year we'll probably be at 90%. I mean, I don't know, but might even be less than that. We've seen similar things in graduate-level math, physics, and biology from models like OpenAi's o1. So if we just continue to extrapolate this in terms of skill that we have, I think if we extrapolate the straight curve, within a few years, we will get to these models being above the highest professional level in terms of humans. Now, will that curve continue? You've pointed to, and I've pointed to a lot of possible reasons why that might not happen. But if the extrapolation curve continues, that is the trajectory we're on.

Lex Fridman
So Anthropic has several competitors. It'd be interesting to get your sort of view of it all. OpenAI, Google, XAI, Meta. What does it take to win in the broad sense of win in this space?

Dario Amodei
Yeah, so I want to separate out a couple things, right? Anthropic's mission is to kind of try to make this all go well. And we have a theory of change called Race to the Top. Race to the Top is about trying to push the other players to do the right thing by setting an example. It's not about being the good guy, it's about setting things up so that all of us can be the good guy. I'll give a few examples of this. Early in the history of Anthropic, one of our co-founders, Chris Olah, who I believe you're interviewing soon, he's the co-founder of the field of mechanistic interpretability, which is an attempt to understand what's going on inside AI models. So we had him and one of our early teams focus on this area of interpretability, which we think is good for making models safe and transparent. For three or four years that had no commercial application whatsoever. It still doesn't. Today we're doing some early betas with it, and probably it will eventually, but this is a very, very long research bed, and one in which we've built in public and shared our results publicly. And we did this because we think it's a way to make models safer. An interesting thing is that as we've done this, other companies have started doing it as well. In some cases because they've been inspired by it, in some cases because they're worried that if other companies are doing this, look more responsible, they want to look more responsible too. No one wants to look like the irresponsible actor. And so they adopt this as well. When folks come to Anthropic, interpretability is often a draw, and I tell them, "The other places you didn't go, tell them why you came here." And then you see soon that there's interpretability teams elsewhere as well. And in a way that takes away our competitive advantage, because it's like, "Oh, now others are doing it as well." But it's good for the broader system, and so we have to invent some new thing that we're doing that others aren't doing as well. And the hope is to basically bid up the importance of doing the right thing. And it's not about us in particular. It's not about having one particular good guy. Other companies can do this as well. If they join the race to do this, that's the best news ever. It's about shaping the incentives to point upward instead of shaping the incentives to point downward.

Lex Fridman
And we should say this example of the field of mechanistic interpretability is just a rigorous non-hand wavy wave doing AI safety-

Dario Amodei
Yes.

Lex Fridman
… or it's tending that way.

Dario Amodei
Trying to. I mean, I think we're still early in terms of our ability to see things, but I've been surprised at how much we've been able to look inside these systems and understand what we see. Unlike with the scaling laws where it feels like there's some law that's driving these models to perform better, on the inside, the models aren't… There's no reason why they should be designed for us to understand them, right? They're designed to operate, they're designed to work. Just like the human brain or human biochemistry. They're not designed for a human to open up the hatch, look inside and understand them. But we have found, and you can talk in much more detail about this to Chris, that when we open them up, when we do look inside them, we find things that are surprisingly interesting.

Lex Fridman
And as a side effect, you also get to see the beauty of these models. You get to explore the beautiful nature of large neural networks through the MEC and TERP kind of methodology.

Dario Amodei
I'm amazed at how clean it's been. I'm amazed at things like induction heads. I'm amazed at things like that we can use sparse auto-encoders to find these directions within the networks, and that the directions correspond to these very clear concepts. We demonstrated this a bit with the Golden Gate Bridge Claude. So this was an experiment where we found a direction inside one of the neural networks layers that corresponded to the Golden Gate Bridge. And we just turned that way up. And so we released this model as a demo, it was kind of half a joke, for a couple days, but it was illustrative of the method we developed. And you could take the model, you could ask it about anything. It would be like you could say, "How was your day?" And anything you asked, because this feature was activated, it would connect to the Golden Gate Bridge. So it would say, I'm feeling relaxed and expansive, much like the arches of the Golden Gate Bridge, or-

Lex Fridman
It would masterfully change topic to the Golden Gate Bridge and integrate it. There was also a sadness to the focus it had on the Golden Gate Bridge. I think people quickly fell in love with it, I think. So people already miss it, because it was taken down, I think after a day.

Dario Amodei
Somehow these interventions on the model, where you kind of adjust its behavior, somehow emotionally made it seem more human than any other version of the model.

Lex Fridman
It's a strong personality, strong identity.

Dario Amodei
It has a strong personality. It has these kind of obsessive interests. We can all think of someone who's obsessed with something. So it does make it feel somehow a bit more human.

Lex Fridman
Let's talk about the present. Let's talk about Claude. So this year, a lot has happened. In March. Claude 3 Opus, Sonnet, Haiku were released. Then Claude 3.5 Sonnet in July, with an updated version just now released. And then also Claude 3.5 Haiku was released. Okay. Can you explain the difference between Opus, Sonnet and Haiku, and how we should think about the different versions?

Dario Amodei
Yeah, so let's go back to March when we first released these three models. So our thinking was different companies produce large and small models, better and worse models. We felt that there was demand, both for a really powerful model, and that might be a little bit slower that you'd have to pay more for, and also for fast cheap models that are as smart as they can be for how fast and cheap. Whenever you want to do some kind of difficult analysis, like if I want to write code for instance, or I want to brainstorm ideas or I want to do creative writing, I want the really powerful model. But then there's a lot of practical applications in a business sense where it's like I'm interacting with a website, I am doing my taxes, or I'm talking to a legal advisor and I want to analyze a contract. Or we have plenty of companies that are just like, I want to do auto-complete on my IDE or something. And for all of those things, you want to act fast and you want to use the model very broadly. So we wanted to serve that whole spectrum of needs. So we ended up with this kind of poetry theme. And so what's a really short poem? It's a haiku. Haiku is the small, fast, cheap model that was at the time, was really surprisingly intelligent for how fast and cheap it was. Sonnet is a medium-sized poem, write a couple paragraphs. And so Sonnet was the middle model. It is smarter but also a little bit slower, a little bit more expensive. And Opus, like a Magnum Opus is a large work, Opus was the largest, smartest model at the time. So that was the original kind of thinking behind it. And our thinking then was, "Well, each new generation of models should shift that trade- off curve." So when we released Sonnet 3.5, it has roughly the same cost and speed as the Sonnet 3 model, but it increased its intelligence to the point where it was smarter than the original Opus 3 model. Especially for code, but also just in general. And so now we've shown results for Haiku 3.5. And I believe Haiku 3.5, the smallest new

model, is about as good as Opus 3, the largest old model. So basically the aim here is to shift the curve and then at some point there's going to be an Opus 3.5. Now every new generation of models has its own thing. They use new data, their personality changes in ways that we try to steer but are not fully able to steer. And so there's never quite that exact equivalence, where the only thing you're changing is intelligence. We always try and improve other things and some things change without us knowing or measuring. So it's very much an inexact science. In many ways, the manner and personality of these models is more an art than it is a science.

Lex Fridman
So what is the reason for the span of time between say, Claude Opus 3.0 and 3.5? What takes that time, if you can speak to it?

Dario Amodei
Yeah, so there's different processes. There's pre-training, which is just kind of the normal language model training. And that takes a very long time. That uses, these days, tens of thousands, sometimes many tens of thousands of GPUs or TPUs or training them, or we use different platforms, but accelerator chips, often training for months. There's then a kind of post-training phase where we do reinforcement learning from human feedback as well as other kinds of reinforcement learning. That phase is getting larger and larger now, and often that's less of an exact science. It often takes effort to get it right. Models are then tested with some of our early partners to see how good they are, and they're then tested, both internally and externally, for their safety, particularly for catastrophic and autonomy risks. So we do internal testing according to our responsible scaling policy, which I could talk more about that in detail. And then we have an agreement with the US and the UK AI Safety Institute, as well as other third-party testers in specific domains, to test the models for what are called CBRN risks, chemical, biological, radiological, and nuclear. We don't think that models pose these risks seriously yet, but every new model we want to evaluate to see if we're starting to get close to some of these more dangerous capabilities. So those are the phases, and then it just takes some time to get the model working in terms of inference and launching it in the API. So there's just a lot of steps to actually making a model work. And of course, we're always trying to make the processes as streamlined as possible. We want our safety testing to be rigorous, but we want it to be rigorous and to be automatic, to happen as fast as it can, without compromising on rigor. Same with our pre-training process and our post-training process. So it's just building anything else. It's just like building airplanes. You want to make them safe, but you want to make the process streamlined. And I think the creative tension between those is an important thing in making the models work.

Lex Fridman
Yeah, rumor on the street, I forget who was saying that, Anthropic has really good tooling. So probably a lot of the challenge here is, on the software engineering side, is to build the tooling to have a efficient, low-friction interaction with the infrastructure.

Dario Amodei
You would be surprised how much of the challenges of building these models comes down to software engineering, performance engineering. From the outside, you might think, "Oh man, we had this Eureka breakthrough." You know, this movie with the science. "We discovered it, we figured it out." But I think all things, even incredible discoveries, they almost always come down to the details. And often super, super boring details. I can't speak to whether we have better tooling than other companies. I mean, haven't been at those other companies, at least not recently, but it's certainly something we give a lot of attention to.

Lex Fridman
I don't know if you can say, but from Claude 3 to Claude 3.5, is there any extra pre-training going on, or is it mostly focused on the post-training? There's been leaps in performance.

Dario Amodei
Yeah, I think at any given stage, we're focused on improving everything at once. Just naturally. Like, there are different teams. Each team makes progress in a particular area, in making their particular segment of the relay race better. And it's just natural that when we make a new model, we put all of these things in at once.

Lex Fridman
So the data you have, the preference data you get from RLHF, is there ways to apply it to newer models as it get trained up?

Dario Amodei
Yeah. Preference data from old models sometimes gets used for new models, although of course it performs somewhat better when it's trained on the new models. Note that we have this constitutional AI method such that we don't only use preference data, there's also a post-training process where we train the model against itself. And there's new types of post-training the model against itself that are used every day. So it's not just RLHF, a bunch of other methods as well. Post-training, I think, is becoming more and more sophisticated.

Lex Fridman
Well, what explains the big leap in performance for the new Sonnet 3.5, I mean, at least in the programming side? And maybe this is a good place to talk about benchmarks. What does it mean to get better? Just the number went up, but I program, but I also love programming, and I Claude 3.5 through Cursor is what I use to assist me in programming.

And there was, at least experientially, anecdotally, it's gotten smarter at programming. So what does it take to get it smarter?

Dario Amodei
We-

Lex Fridman
So what does it take to get it smarter?

Dario Amodei
We observe that as well. By the way, there were a couple very strong engineers here at Anthropic, who all previous code models, both produced by us and produced by all the other companies, hadn't really been useful to them. They said, "Maybe this is useful to a beginner. It's not useful to me." But Sonnet 3.5, the original one for the first time, they said, "Oh, my God, this helped me with something that it would've taken me hours to do. This is the first model that's actually saved me time." So again, the water line is rising. And then I think the new Sonnet has been even better. In terms of what it takes, I'll just say it's been across the board. It's in the pre-training, it's in the post-training, it's in various evaluations that we do. We've observed this as well. And if we go into the details of the benchmark, so SWE-bench is basically… Since you're a programmer, you'll be familiar with pull requests, and just pull requests, they're like a sort of atomic unit of work. You could say I'm implementing one thing. So SWE-bench actually gives you a real world situation where the code base is in a current state and I'm trying to implement something that's described in language. We have internal benchmarks where we measure the same thing and you say, "Just give the model free rein to do anything, run anything, edit anything. How well is it able to complete these tasks?" And it's that benchmark that's gone from "it can do it 3% of the time" to "it can do it about 50% of the time." So I actually do believe that you can gain benchmarks, but I think if we get to 100% on that benchmark in a way that isn't over-trained or game for that particular benchmark, probably represents a real and serious increase in programming ability. And I would suspect that if we can get to 90-95% that it will represent ability to autonomously do a significant fraction of software engineering tasks.

Lex Fridman
Well, ridiculous timeline question. When is Claude Opus 3.5 coming up?

Dario Amodei
Not giving you an exact date, but as far as we know, the plan is still to have a Claude 3.5 Opus.

Lex Fridman
Are we going to get it before GTA 6 or no?

Dario Amodei
Like Duke Nukem Forever?

Lex Fridman
Duke Nukem. Right.

Dario Amodei
What was that game? There was some game that was delayed 15 years.

Lex Fridman
That's right.

Dario Amodei
Was that Duke Nukem Forever?

Lex Fridman
Yeah. And I think GTA is now just releasing trailers.

Dario Amodei
It's only been three months since we released the first Sonnet.

Lex Fridman
Yeah, it's the incredible pace of release.

Dario Amodei
It just tells you about the pace, the expectations for when things are going to come out.

Lex Fridman
So what about 4.0? So how do you think, as these models get bigger and bigger, about versioning and also just versioning in general, why Sonnet 3.5 updated with the date? Why not Sonnet 3.6, which a lot of people are calling it?

Dario Amodei
Naming is actually an interesting challenge here, right? Because I think a year ago, most of the model was pre-training. And so you could start from the beginning and just say, "Okay, we're going to have models of different sizes. We're going to train them all together and we'll have a family of naming schemes and then we'll put some new magic into them and then we'll have the next generation." The trouble starts already when some of them take a lot longer than others to train. That already messes up your time a little bit. But as you make big improvement in pre-training, then you suddenly notice, "Oh, I can make better pre-train model." And that doesn't take very long to do, but clearly it has the same size and shape of previous models. So I think those two together as well as the timing issues. Any kind of

scheme you come up with, the reality tends to frustrate that scheme, right? It tends to break out of the scheme. It's not like software where you can say, "Oh, this is 3.7, this is 3.8." No, you have models with different trade-offs. You can change some things in your models, you can change other things. Some are faster and slower at inference. Some have to be more expensive, some have to be less expensive. And so I think all the companies have struggled with this. I think we were in a good position in terms of naming when we had Haiku, Sonnet and Opus.

Lex Fridman
It was great, great start.

Dario Amodei
We're trying to maintain it, but it's not perfect, so we'll try and get back to the simplicity. But just the nature of the field, I feel like no one's figured out naming. It's somehow a different paradigm from normal software and so none of the companies have been perfect at it. It's something we struggle with surprisingly much relative to how trivial it is for the grand science of training the models.

Lex Fridman
So from the user side, the user experience of the updated Sonnet 3.5 is just different than the previous June 2024 Sonnet 3.5. It would be nice to come up with some kind of labeling that embodies that. Because people talk about Sonnet 3.5, but now there's a different one. And so how do you refer to the previous one and the new one when there's a distinct improvement? It just makes conversation about it just challenging.

Dario Amodei
Yeah, yeah. I definitely think this question of there are lots of properties of the models that are not reflected in the benchmarks. I think that's definitely the case and everyone agrees. And not all of them are capabilities. Models can be polite or brusque, they can be very reactive or they can ask you questions. They can have what feels like a warm personality or a cold personality. They can be boring or they can be very distinctive like Golden Gate Claude was. And we have a whole team focused on, I think we call it Claude character. Amanda leads that team and we'll talk to you about that, but it's still a very inexact science and often we find that models have properties that we're not aware of. The fact of the matter is that you can talk to a model 10,000 times and there are some behaviors you might not see just like with a human, right? I can know someone for a few months and not know that they have a certain skill or not know that there's a certain side to them. And so I think we just have to get used to this idea. And we're always looking for better ways of testing our models to demonstrate these capabilities and also to decide which are the personality properties we want models to have and which we don't want to have. That itself, the normative question, is also super interesting.

Lex Fridman
I got to ask you a question from Reddit.

Dario Amodei
From Reddit? Oh, boy.

Lex Fridman
There's just this fascinating, to me at least, it's a psychological social phenomenon where people report that Claude has gotten dumber for them over time. And so the question is, does the user complaint about the dumbing down of Claude 3.5 Sonnet hold any water? So are these anecdotal reports a kind of social phenomena or is there any cases where Claude would get dumber?

Dario Amodei
So this actually doesn't apply. This isn't just about Claude. I believe I've seen these complaints for every foundation model produced by a major company. People said this about GPT-4, they said it about GPT-4 Turbo. So a couple things. One, the actual weights of the model, the actual brain of the model, that does not change unless we introduce a new model. There are just a number of reasons why it would not make sense practically to be randomly substituting in new versions of the model. It's difficult from an inference perspective and it's actually hard to control all the consequences of changing the weights of the model. Let's say you wanted to fine-tune the model, I don't know, to say "certainly" less, which an old version of Sonnet used to do. You actually end up changing 100 things as well. So we have a whole process for it and we have a whole process for modifying the model. We do a bunch of testing on it. We do a bunch of user testing in early customers. So we both have never changed the weights of the model without telling anyone. And certainly, in the current setup, it would not make sense to do that. Now, there are a couple things that we do occasionally do. One is sometimes we run A/B tests, but those are typically very close to when a model is being released and for a very small fraction of time. So the day before the new Sonnet 3.5, I agree we should have had a better name. It's clunky to refer to it. There were some comments from people that it's gotten a lot better and that's because a fraction we're exposed to an A/B test for those one or two days. The other is that occasionally the system prompt will change. The system prompt can have some effects, although it's unlikely to dumb down models, it's unlikely to make them dumber. And we've seen that while these two things, which I'm listing to be very complete, happened quite infrequently, the complaints for us and for other model companies about the model change, the model isn't good at this, the model got more censored, the model was dumbed down. Those complaints are constant and so I don't want to say people are imagining it or anything, but the models are, for the most part, not changing. If I were to offer a theory, I think it actually relates to one of the things I said before, which is that models are very complex and have many aspects to them. And so often, if I ask the model a question, if I'm like, "Do task X" versus, "Can you do task X?" the model might respond in different ways. And

so there are all kinds of subtle things that you can change about the way you interact with the model that can give you very different results. To be clear, this itself is like a failing by us and by the other model providers that the models are just often sensitive to small changes in wording. It's yet another way in which the science of how these models work is very poorly developed. And so if I go to sleep one night and I was talking to the model in a certain way and I slightly changed the phrasing of how I talk to the model, I could get different results. So that's one possible way. The other thing is, man, it's just hard to quantify this stuff. It's hard to quantify this stuff. I think people are very excited by new models when they come out and then as time goes on, they become very aware of their limitations. So that may be another effect, but that's all a very long-winded way of saying for the most part, with some fairly narrow exceptions, the models are not changing.

Lex Fridman
I think there is a psychological effect. You just start getting used to it, the baseline raises. When people who have first gotten Wi-Fi on airplanes, it's amazing, magic.

Dario Amodei
It's amazing. Yeah.

Lex Fridman
And then you start-

Dario Amodei
And now I'm like, "I can't get this thing to work. This is such a piece of crap."

Lex Fridman
Exactly. So it's easy to have the conspiracy theory of, "They're making Wi-Fi slower and slower." This is probably something I'll talk to Amanda much more about, but another Reddit question, "When will Claude stop trying to be my pure tentacle grandmother imposing its moral worldview on me as a paying customer? And also, what is the psychology behind making Claude overly apologetic?" So this reports about the experience, a different angle on the frustration. It has to do with the character [inaudible 00:47:06].

Dario Amodei
Yeah, so a couple points on this first. One is things that people say on Reddit and Twitter or X or whatever it is, there's actually a huge distribution shift between the stuff that people complain loudly about on social media and what actually statistically users care about and that drives people to use the models. People are frustrated with things like the model not writing out all the code or the model just not being as good at code as it could be, even though it's the best model in the world on code. I think the majority of things are about that, but certainly a vocal minority raise these concerns, are frustrated by the model refusing things that it shouldn't refuse or apologizing too much or just having these annoying verbal

tics. The second caveat, and I just want to say this super clearly because I think some people don't know it, others know it, but forget it. It is very difficult to control across the board how the models behave. You cannot just reach in there and say, "Oh, I want the model to apologize less." You can do that. You can include training data that says, "Oh, the model should apologize less." But then in some other situation, they end up being super rude or overconfident in a way that's misleading people. So there are all these trade-offs. For example, another thing is if there was a period during which models, ours and I think others as well, were too verbose, they would repeat themselves, they would say too much. You can cut down on the verbosity by penalizing the models for just talking for too long. What happens when you do that, if you do it in a crude way, is when the models are coding, sometimes they'll say, "Rest of the code goes here," right? Because they've learned that that's the way to economize and that they see it. And then so that leads the model to be so-called lazy in coding where they're just like, "Ah, you can finish the rest of it." It's not because we want to save on compute or because the models are lazy during winter break or any of the other conspiracy theories that have come up. Actually, it's just very hard to control the behavior of the model, to steer the behavior of the model in all circumstances at once. There's this whack- a-mole aspect where you push on one thing and these other things start to move as well that you may not even notice or measure. And so one of the reasons that I care so much about grand alignment of these AI systems in the future is actually, these systems are actually quite unpredictable. They're actually quite hard to steer and control. And this version we're seeing today of you make one thing better, it makes another thing worse, I think that's like a present day analog of future control problems in AI systems that we can start to study today. I think that difficulty in steering the behavior and making sure that if we push an AI system in one direction, it doesn't push it in another direction in some other ways that we didn't want. I think that's an early sign of things to come, and if we can do a good job of solving this problem of you ask the model to make and distribute smallpox and it says no, but it's willing to help you in your graduate level virology class, how do we get both of those things at once? It's hard. It's very easy to go to one side or the other and it's a multidimensional problem. And so I think these questions of shaping the model's personality, I think they're very hard. I think we haven't done perfectly on them. I think we've actually done the best of all the AI companies, but still so far from perfect. And I think if we can get this right, if we can control the false positives and false negatives in this very controlled present day environment, we'll be much better at doing it for the future when our worry is: will the models be super autonomous? Will they be able to make very dangerous things? Will they be able to autonomously build whole companies and are those companies aligned? So I think of this present task as both vexing but also good practice for the future.

Lex Fridman
What's the current best way of gathering user feedback? Not anecdotal data, but just large-scale data about pain points or the opposite of pain points, positive things, so on? Is it internal testing? Is it a specific group testing, A/B testing? What works?

Dario Amodei
So typically, we'll have internal model bashings where all of Anthropic… Anthropic is almost 1,000 people. People just try and break the model. They try and interact with it various ways. We have a suite of evals for, "Oh, is the model refusing in ways that it couldn't?" I think we even had a "certainly" eval because again, at one point, the model had this problem where it had this annoying tick where it would respond to a wide range of questions by saying, "Certainly, I can help you with that. Certainly, I would be happy to do that. Certainly, this is correct." And so we had a "certainly" eval, which is: how often does the model say certainly? But look, this is just a whack-a-mole. What if it switches from "certainly" to "definitely"? So every time we add a new eval and we're always evaluating for all the old things, we have hundreds of these evaluations, but we find that there's no substitute for a human interacting with it. And so it's very much like the ordinary product development process. We have hundreds of people within Anthropic bash the model. Then we do external A/B tests. Sometimes we'll run tests with contractors. We pay contractors to interact with the model. So you put all of these things together and it's still not perfect. You still see behaviors that you don't quite want to see. You still see the model refusing things that it just doesn't make sense to refuse. But I think trying to solve this challenge, trying to stop the model from doing genuinely bad things that everyone agrees it shouldn't do, everyone agrees that the model shouldn't talk about, I don't know, child abuse material. Everyone agrees the model shouldn't do that, but at the same time, that it doesn't refuse in these dumb and stupid ways. I think drawing that line as finely as possible, approaching perfectly, is still a challenge and we're getting better at it every day, but there's a lot to be solved. And again, I would point to that as an indicator of a challenge ahead in terms of steering much more powerful models.

Lex Fridman
Do you think Claude 4.0 is ever coming out?

Dario Amodei
I don't want to commit to any naming scheme because if I say here, "We're going to have Claude 4 next year," and then we decide that we should start over because there's a new type of model, I don't want to commit to it. I would expect in a normal course of business that Claude 4 would come after Claude 3. 5, but you never know in this wacky field.

Lex Fridman
But this idea of scaling is continuing.

Dario Amodei
Scaling is continuing. There will definitely be more powerful models coming from us than the models that exist today. That is certain. Or if there aren't, we've deeply failed as a company.

Lex Fridman
Okay. Can you explain the responsible scaling policy and the AI safety level standards, ASL levels?

Dario Amodei
As much as I am excited about the benefits of these models, and we'll talk about that if we talk about Machines of Loving Grace, I'm worried about the risks and I continue to be worried about the risks. No one should think that Machines of Loving Grace was me saying I'm no longer worried about the risks of these models. I think they're two sides of the same coin. The power of the models and their ability to solve all these problems in biology, neuroscience, economic development, governance and peace, large parts of the economy, those come with risks as well, right? With great power comes great responsibility. The two are paired. Things that are powerful can do good things and they can do bad things. I think of those risks as being in several different categories, perhaps the two biggest risks that I think about. And that's not to say that there aren't risks today that are important, but when I think of really the things that would happen on the grandest scale, one is what I call catastrophic misuse. These are misuse of the models in domains like cyber, bio, radiological, nuclear, things that could harm or even kill thousands, even millions of people if they really, really go wrong. These are the number one priority to prevent. And here I would just make a simple observation, which is that the models, if I look today at people who have done really bad things in the world, I think actually humanity has been protected by the fact that the overlap between really smart, well-educated people and people who want to do really horrific things has generally been small. Let's say I'm someone who I have a PhD in this field, I have a well-paying job. There's so much to lose. Even assuming I'm completely evil, which most people are not, why would such a person risk their life, risk their legacy, their reputation to do something truly, truly evil? If we had a lot more people like that, the world would be a much more dangerous place. And so my worry is that by being a much more intelligent agent, AI could break that correlation. And so I do have serious worries about that. I believe we can prevent those worries. But I think as a counterpoint to Machines of Loving Grace, I want to say that there's still serious risks. And the second range of risks would be the autonomy risks, which is the idea that models might, on their own, particularly as we give them more agency than they've had in the past, particularly as we give them supervision over wider tasks like writing whole code bases or someday even effectively operating entire companies, they're on a long enough leash. Are they doing what we really want them to do? It's very difficult to even understand in detail what they're doing, let alone control it. And like I said, these early signs that it's hard to perfectly draw the boundary between things the model should do and things the model shouldn't do that if you go to one side, you get things that are annoying and useless and you go to the other side, you get other behaviors. If you fix one thing, it creates other problems. We're getting better and better at solving this. I don't think this is an unsolvable problem. I think this is a science like the safety of airplanes or the safety of cars or the safety of drugs. I don't think there's any big thing we're missing. I just think we need to get better at controlling these models.

And so these are the two risks I'm worried about. And our responsible scaling plan, which I'll recognize is a very long-winded answer to your question.

Lex Fridman
I love it. I love it.

Dario Amodei
Our responsible scaling plan is designed to address these two types of risks. And so every time we develop a new model, we basically test it for its ability to do both of these bad things. So if I were to back up a little bit, I think we have an interesting dilemma with AI systems where they're not yet powerful enough to present these catastrophes. I don't know if they'll ever present these catastrophes. It's possible they won't. But the case for worry, the case for risk is strong enough that we should act now and they're getting better very, very fast. I testified in the Senate that we might have serious bio risks within two to three years. That was about a year ago. Things have proceeded apace. So we have this thing where it's surprisingly hard to address these risks because they're not here today, they don't exist. They're like ghosts, but they're coming at us so fast because the models are improving so fast. So how do you deal with something that's not here today, doesn't exist, but is coming at us very fast? So the solution we came up with for that, in collaboration with people like the organization METR and Paul Christiano is what you need for that are you need tests to tell you when the risk is getting close. You need an early warning system. And so every time we have a new model, we test it for its capability to do these CBRN tasks as well as testing it for how capable it is of doing tasks autonomously on its own. And in the latest version of our RSP, which we released in the last month or two, the way we test autonomy risks is the AI model's ability to do aspects of AI research itself, which when the AI models can do AI research, they become truly, truly autonomous. And that threshold is important for a bunch of other ways. And so what do we then do with these tasks? The RSP basically develops what we've called an if-then structure, which is if the models pass a certain capability, then we impose a certain set of safety and security requirements on them. So today's models are what's called ASL-2. Models that were ASL-1 is for systems that manifestly don't pose any risk of autonomy or misuse. So for example, a chess playing bot, Deep Blue would be ASL-1. It's just manifestly the case that you can't use Deep Blue for anything other than chess. It was just designed for chess. No one's going to use it to conduct a masterful cyber attack or to run wild and take over the world. ASL-2 is today's AI systems where we've measured them and we think these systems are simply not smart enough to autonomously self-replicate or conduct a bunch of tasks and also not smart enough to provide meaningful information about CBRN risks and how to build CBRN weapons above and beyond what can be known from looking at Google. In fact, sometimes they do provide information above and beyond a search engine, but not in a way that can be stitched together, not in a way that end-to-end is dangerous enough. So ASL-3 is going to be the point at which the models are helpful enough to enhance the capabilities of non-state actors, right? State actors can already do, unfortunately, to a high level of

proficiency, a lot of these very dangerous and destructive things. The difference is that non-state actors are not capable of it. And so when we get to ASL-3, we'll take special security precautions designed to be sufficient to prevent theft of the model by non-state actors and misuse of the model as it's deployed. We'll have to have enhanced filters targeted at these particular areas.

Lex Fridman
Cyber, bio, nuclear.

Dario Amodei
Cyber, bio, nuclear and model autonomy, which is less a misuse risk and more a risk of the model doing bad things itself. ASL-4, getting to the point where these models could enhance the capability of a already knowledgeable state actor and/or become the main source of such a risk. If you wanted to engage in such a risk, the main way you would do it is through a model. And then I think ASL-4 on the autonomy side, it's some amount of acceleration in AI research capabilities with an AI model. And then ASL-5 is where we would get to the models that are truly capable that it could exceed humanity in their ability to do any of these tasks. And so the point of the if-then structure commitment is basically to say, "Look, I don't know, I've been working with these models for many years and I've been worried about risk for many years. It's actually dangerous to cry wolf. It's actually dangerous to say this model is risky. And people look at it and they say this is manifestly not dangerous." Again, it's the delicacy of the risk isn't here today, but it's coming at us fast. How do you deal with that? It's really vexing to a risk planner to deal with it. And so this if-then structure basically says, "Look, we don't want to antagonize a bunch of people, we don't want to harm our own ability to have a place in the conversation by imposing these very onerous burdens on models that are not dangerous today." So the if-then, the trigger commitment is basically a way to deal with this. It says you clamp down hard when you can show the model is dangerous. And of course, what has to come with that is enough of a buffer threshold that you're not at high risk of missing the danger. It's not a perfect framework. We've had to change it. We came out with a new one just a few weeks ago and probably going forward, we might release new ones multiple times a year because it's hard to get these policies right technically, organizationally from a research perspective. But that is the proposal, if-then commitments and triggers in order to minimize burdens and false alarms now, but really react appropriately when the dangers are here.

Lex Fridman
What do you think the timeline for ASL-3 is where several of the triggers are fired? And what do you think the timeline is for ASL-4?

Dario Amodei
Yeah. So that is hotly debated within the company. We are working actively to prepare ASL-3 security measures as well as ASL-3 deployment measures. I'm not going to go into

detail, but we've made a lot of progress on both and we're prepared to be, I think, ready quite soon. I would not be surprised at all if we hit ASL-3 next year. There was some concern that we might even hit it this year. That's still possible. That could still happen. It's very hard to say, but I would be very, very surprised if it was 2030. I think it's much sooner than that.

Lex Fridman
So there's protocols for detecting it, the if-then and then there's protocols for how to respond to it.

Dario Amodei
Yes.

Lex Fridman
How difficult is the second, the latter?

Dario Amodei
Yeah. I think for ASL-3, it's primarily about security and about filters on the model relating to a very narrow set of areas when we deploy the model. Because at ASL-3, the model isn't autonomous yet. And so you don't have to worry about the model itself behaving in a bad way even when it's deployed internally. So I think the ASL- 3 measures are, I won't say straightforward, they're rigorous, but they're easier to reason about. I think once we get to ASL-4, we start to have worries about the models being smart enough that they might sandbag tests, they might not tell the truth about tests. We had some results came out about sleeper agents and there was a more recent paper about, "Can the models mislead attempts to sandbag their own abilities, present themselves as being less capable than they are?" And so I think with ASL-4, there's going to be an important component of using other things than just interacting with the models. For example, interpretability or hidden chains of thought where you have to look inside the model and verify via some other mechanism that is not as easily corrupted as what the model says, that the model indeed has some property. So we're still working on ASL-4. One of the properties of the RSP is that we don't specify ASL-4 until we've hit ASL-3. And I think that's proven to be a wise decision because even with ASL-3, again, it's hard to know this stuff in detail, and we want to take as much time as we can possibly take to get these things right.

Lex Fridman
So for ASL-3, the bad actor will be the humans.

Dario Amodei
Humans, yes.

Lex Fridman
And so there's a little bit more…

Dario Amodei
For ASL- 4, it's both, I think.

Lex Fridman
It's both. And so deception, and that's where mechanistic interpretability comes into play, and hopefully the techniques used for that are not made accessible to the model.

Dario Amodei
Yeah. Of course, you can hook up the mechanistic interpretability to the model itself, but then you've lost it as a reliable indicator of the model state. There are a bunch of exotic ways you can think of that it might also not be reliable, like if the model gets smart enough that it can jump computers and read the code where you're looking at its internal state. We've thought about some of those. I think they're exotic enough. There are ways to render them unlikely. But yeah, generally, you want to preserve mechanistic interpretability as a verification set or test set that's separate from the training process of the model.

Lex Fridman
See, I think as these models become better and better conversation and become smarter, social engineer becomes a threat too because they could start being very convincing to the engineers inside companies.

Dario Amodei
Oh, yeah. Yeah. We've seen lots of examples of demagoguery in our life from humans, and there's a concern that models could do that as well.

Lex Fridman
One of the ways that Claude has been getting more and more powerful is it's now able to do some agentic stuff, computer use. There's also an analysis within the sandbox of Claude.ai itself. But let's talk about computer use. That seems to me super exciting that you can just give Claude a task and it takes a bunch of actions, figures it out, and has access to the… … a bunch of actions, figures it out and has access to your computer through screenshots. So can you explain how that works and where that's headed?

Dario Amodei
Yeah. It's actually relatively simple. So Claude has had for a long time, since Claude 3 back in March, the ability to analyze images and respond to them with text. The only new thing we added is those images can be screenshots of a computer and in response, we train the model to give a location on the screen where you can click and/or buttons on the keyboard, you can press in order to take action. And it turns out that with actually not all that much additional training, the models can get quite good at that task. It's a good example of generalization. People sometimes say if you get to lower earth orbit, you're halfway to anywhere because of how much it takes to escape the gravity well. If you have a strong

pre-trained model, I feel like you're halfway to anywhere in terms of the intelligence space. And so actually, it didn't take all that much to get Claude to do this. And you can just set that in a loop, give the model a screenshot, tell it what to click on, give it the next screenshot, tell it what to click on and that turns into a full kind of almost 3D video interaction of the model and it's able to do all of these tasks. We showed these demos where it's able to fill out spreadsheets, it's able to kind of interact with a website, it's able to open all kinds of programs, different operating systems, Windows, Linux, Mac. So I think all of that is very exciting. I will say, while in theory there's nothing you could do there that you couldn't have done through just giving the model the API to drive the computer screen, this really lowers the barrier. And there's a lot of folks who either aren't in a position to interact with those APIs or it takes them a long time to do. It's just the screen is just a universal interface that's a lot easier to interact with. And so I expect over time, this is going to lower a bunch of barriers. Now, honestly, the current model has, it leaves a lot still to be desired and we were honest about that in the blog. It makes mistakes, it misclicks. We were careful to warn people, "Hey, you can't just leave this thing to run on your computer for minutes and minutes. You got to give this thing boundaries and guardrails." And I think that's one of the reasons we released it first in an API form rather than just hand the consumer and give it control of their computer. But I definitely feel that it's important to get these capabilities out there. As models get more powerful, we're going to have to grapple with how do we use these capabilities safely. How do we prevent them from being abused? And I think releasing the model while the capabilities are still limited is very helpful in terms of doing that. I think since it's been released, a number of customers, I think Replit was maybe one of the most quickest to deploy things, have made use of it in various ways. People have hooked up demos for Windows desktops, Macs, Linux machines. So yeah, it's been very exciting. I think as with anything else, it comes with new exciting abilities and then with those new exciting abilities, we have to think about how to make the model safe, reliable, do what humans want them to do. It's the same story for everything. Same thing. It's that same tension.

Lex Fridman
But the possibility of use cases here, just the range is incredible. So how much to make it work really well in the future? How much do you have to specially kind of go beyond what the pre-trained model is doing, do more post-training, RLHF or supervised fine-tuning or synthetic data just for the agentive stuff?

Dario Amodei
Yeah. I think speaking at a high level, it's our intention to keep investing a lot in making the model better. I think we look at some of the benchmarks where previous models were like, "Oh, could do it 6% of the time," and now our model would do it 14 or 22% of the time. And yeah, we want to get up to the human level reliability of 80, 90% just like anywhere else. We're on the same curve that we were on with SWE-bench where I think I would guess a year from now, the models can do this very, very reliably. But you gotta start somewhere.

Lex Fridman
So you think it's possible to get to the human level 90% basically doing the same thing you're doing now or it has to be special for computer use?

Dario Amodei
It depends what you mean by special and special in general, but I generally think the same kinds of techniques that we've been using to train the current model, I expect that doubling down on those techniques in the same way that we have for code, for models in general, for image input, for voice, I expect those same techniques will scale here as they have everywhere else,

Lex Fridman
But this is giving the power of action to Claude and so you could do a lot of really powerful things, but you could do a lot of damage also.

Dario Amodei
Yeah, yeah. No and we've been very aware of that. Look, my view actually is computer use isn't a fundamentally new capability like the CBRN or autonomy capabilities are. It's more like it kind of opens the aperture for the model to use and apply its existing abilities. And so the way we think about it, going back to our RSP, is nothing that this model is doing inherently increases the risk from an RSP perspective, but as the models get more powerful, having this capability may make it scarier once it has the cognitive capability to do something at the ASL-3 and ASL-4 level, this may be the thing that kind of unbounds it from doing so. So going forward, certainly this modality of interaction is something we have tested for and that we will continue to test for an RSP going forward. I think it's probably better to learn and explore this capability before the model is super capable

Lex Fridman
Yeah. And there's a lot of interesting attacks like prompt injection because now you've widened the aperture so you can prompt inject through stuff on screen. So if this becomes more and more useful, then there's more and more benefit to inject stuff into the model. If it goes to certain web page, it could be harmless stuff like advertisements or it could be harmful stuff, right?

Dario Amodei
Yeah, we've thought a lot about things like spam, CAPTCHA, mass… One secret, I'll tell you, if you've invented a new technology, not necessarily the biggest misuse, but the first misuse you'll see, scams, just petty scams.

Lex Fridman
Yeah.

Dario Amodei
It's like a thing as old, people scamming each other, it's this thing as old as time. And it's just every time, you got to deal with it.

Lex Fridman
It's almost silly to say, but it's true, sort of bots and spam in general is a thing as it gets more and more intelligent-

Dario Amodei
Yeah, yeah.

Lex Fridman
… it's harder and harder to fight it.

Dario Amodei
Like I said, there are a lot of petty criminals in the world and it's like every new technology is a new way for petty criminals to do something stupid and malicious.

Lex Fridman
Is there any ideas about sandboxing it? How difficult is the sandboxing task?

Dario Amodei
Yeah, we sandbox during training. So for example, during training we didn't expose the model to the internet. I think that's probably a bad idea during training because the model can be changing its policy, it can be changing what it's doing and it's having an effect in the real world. In terms of actually deploying the model, it kind of depends on the application. Sometimes you want the model to do something in the real world. But of course, you can always put guard, you can always put guard rails on the outside. You can say, "Okay, well, this model's not going to move data from my, the model's not going to move any files from my computer or my web server to anywhere else." Now, when you talk about sandboxing, again, when we get to ASL-4, none of these precautions are going to make sense there. When you talk about ASL-4, you're then, the model is being, there's theoretical worry the model could be smart enough to kind of break it to out of any box. And so there, we need to think about mechanistic interpretability. If we're going to have a sandbox, it would need to be a mathematically provable. That's a whole different world than what we're dealing with with the models today.

Lex Fridman
Yeah, the science of building a box from which ASL-4 AI system cannot escape.

Dario Amodei
I think it's probably not the right approach. I think the right approach, instead of having something unaligned that you're trying to prevent it from escaping, I think it's better to just design the model the right way or have a loop where you look inside the model and you're able to verify properties and that gives you an opportunity to tell, iterate and actually get it right. I think containing bad models is a much worse solution than having good models.

Lex Fridman
Let me ask about regulation. What's the role of regulation in keeping AI safe? So for example, can you describe California AI regulation bill SB 1047 that was ultimately vetoed by the governor? What are the pros and cons of this bill in general?

Dario Amodei
Yes, we ended up making some suggestions to the bill. And then some of those were adopted and we felt, I think, quite positively about the bill by the end of that, it did still have some downsides. And of course, it got vetoed. I think at a high level, I think some of the key ideas behind the bill are I would say similar to ideas behind our RSPs. And I think it's very important that some jurisdiction, whether it's California or the federal government and/or other countries and other states, passes some regulation like this. And I can talk through why I think that's so important. So I feel good about our RSP. It's not perfect. It needs to be iterated on a lot. But it's been a good forcing function for getting the company to take these risks seriously, to put them into product planning, to really make them a central part of work at Anthropic and to make sure that all of a thousand people, and it's almost a thousand people now at Anthropic, understand that this is one of the highest priorities of the company, if not the highest priority. But one, there are still some companies that don't have RSP like mechanisms, like OpenAI, Google did adopt these mechanisms a couple months after Anthropic did, but there are other companies out there that don't have these mechanisms at all. And so if some companies adopt these mechanisms and others don't, it's really going to create a situation where some of these dangers have the property that it doesn't matter if three out of five of the companies are being safe, if the other two are being unsafe, it creates this negative externality. And I think the lack of uniformity is not fair to those of us who have put a lot of effort into being very thoughtful about these procedures. The second thing is I don't think you can trust these companies to adhere to these voluntary plans on their own. Right? I like to think that Anthropic will, we do everything we can that we will, our RSP is checked by our long-term benefit trust, so we do everything we can to adhere to our own RSP. But you hear lots of things about various companies saying, "Oh, they said they would give this much compute and they didn't. They said they would do this thing and the didn't." I don't think it makes sense to litigate particular things that companies have done, but I think this broad principle that if there's nothing watching over them, if there's nothing watching over us as an industry, there's no guarantee that we'll do the right thing and the stakes are very high. And so I think it's important to have a uniform standard that everyone follows and to make sure that simply that the industry does what a majority of

the industry has already said is important and has already said that they definitely will do. Right, some people, I think there's a class of people who are against regulation on principle. I understand where that comes from. If you go to Europe and you see something like GDPR, you see some of the other stuff that they've done. Some of it's good, but some of it is really unnecessarily burdensome and I think it's fair to say really has slowed innovation. And so I understand where people are coming from on priors. I understand why people start from that position. But again, I think AI is different. If we go to the very serious risks of autonomy and misuse that I talked about just a few minutes ago, I think that those are unusual and they warrant an unusually strong response. And so I think it's very important. Again, we need something that everyone can get behind. I think one of the issues with SB 1047, especially the original version of it was it had a bunch of the structure of RSPs, but it also had a bunch of stuff that was either clunky or that just would've created a bunch of burdens, a bunch of hassle and might even have missed the target in terms of addressing the risks. You don't really hear about it on Twitter, you just hear about kind of people are cheering for any regulation. And then the folks who are against make up these often quite intellectually dishonest arguments about how it'll make us move away from California, bill doesn't apply if you're headquartered in California, bill only applies if you do business in California, or that it would damage the open source ecosystem or that it would cause all of these things. I think those were mostly nonsense, but there are better arguments against regulation. There's one guy, Dean Ball, who's really, I think, a very scholarly analyst who looks at what happens when a regulation is put in place in ways that they can kind of get a life of their own or how they can be poorly designed. And so our interest has always been we do think there should be regulation in this space, but we want to be an actor who makes sure that that regulation is something that's surgical, that's targeted at the serious risks and is something people can actually comply with. Because something I think the advocates of regulation don't understand as well as they could is if we get something in place that's poorly targeted, that wastes a bunch of people's time, what's going to happen is people are going to say, "See, these safety risks, this is nonsense. I just had to hire 10 lawyers to fill out all these forms. I had to run all these tests for something that was clearly not dangerous." And after six months of that, there will be a ground swell and we'll end up with a durable consensus against regulation. And so I think the worst enemy of those who want real accountability is badly designed regulation. We need to actually get it right. And if there's one thing I could say to the advocates, it would be that I want them to understand this dynamic better and we need to be really careful and we need to talk to people who actually have experience seeing how regulations play out in practice. And the people who have seen that, understand to be very careful. If this was some lesser issue, I might be against regulation at all. But what I want the opponents to understand is that the underlying issues are actually serious. They're not something that I or the other companies are just making up because of regulatory capture, they're not sci-fi fantasies, they're not any of these things. Every time we have a new model, every few months we measure the behavior of these models and they're getting better and better at these concerning tasks just as they are getting better and better at good, valuable, economically useful tasks. And so I would just love it if some of the former, I think SB 1047 was very polarizing, I would love it if some of the most reasonable opponents and some of the most reasonable proponents would sit down together. And I think that the different AI companies, Anthropic was the only AI company that felt positively in a very detailed way. I think Elon tweeted briefly something positive, but some of the big ones like Google, OpenAI, Meta, Microsoft were pretty staunchly against. So I would really is if some of the key stakeholders, some of the most thoughtful proponents and some of the most thoughtful opponents would sit down and say how do we solve this problem in a way that the proponents feel brings a real reduction in risk and that the opponents feel that it is not hampering the industry or hampering innovation any more necessary than it needs to. I think for whatever reason, that things got too polarized and those two groups didn't get to sit down in the way that they should. And I feel urgency. I really think we need to do something in 2025. If we get to the end of 2025 and we've still done nothing about this, then I'm going to be worried. I'm not worried yet because, again, the risks aren't here yet, but I think time is running short.

Lex Fridman
And come up with something surgical, like you said.

Dario Amodei
Yeah, yeah, yeah, exactly. And we need to get away from this intense pro safety versus intense anti-regulatory rhetoric. It's turned into these flame wars on Twitter and nothing good's going to come of that.

Lex Fridman
So there's a lot of curiosity about the different players in the game. One of the OGs is OpenAI. You've had several years of experience at OpenAI. What's your story and history there?

Dario Amodei
Yeah. So I was at OpenAI for roughly five years. For the last, I think it was couple years, I was vice president of research there. Probably myself and Ilya Sutskever were the ones who really kind of set the research direction. Around 2016 or 2017, I first started to really believe in or at least confirm my belief in the scaling hypothesis when Ilya famously said to me, "The thing you need to understand about these models is they just want to learn. The models just want to learn." And again, sometimes there are these one sentences, these then cones, that you hear them and you're like, "Ah, that explains everything. That explains a thousand things that I've seen." And then ever after, I had this visualization in my head of you optimize the models in the right way, you point the models in the right way, they just want to learn. They just want to solve the problem regardless of what the problem is.

Lex Fridman
So get out of their way, basically?

Dario Amodei
Get out of their way. Yeah.

Lex Fridman
Okay.

Dario Amodei
Don't impose your own ideas about how they should learn. And this was the same thing as Rich Sutton put out in the bitter lesson or Gwern put out in the scaling hypothesis. I think generally the dynamic was I got this kind of inspiration from Ilya and from others, folks like Alec Radford, who did the original GPT-1 and then ran really hard with it, me and my collaborators, on GPT-2, GPT-3, RL from Human Feedback, which was an attempt to kind of deal with the early safety and durability, things like debate and amplification, heavy on interpretability. So again, the combination of safety plus scaling. Probably 2018, 2019, 2020, those were kind of the years when myself and my collaborators, probably many of whom became co-founders of Anthropic, kind of really had a vision and drove the direction.

Lex Fridman
Why'd you leave? Why'd you decide to leave?

Dario Amodei
Yeah, so look, I'm going to put things this way and I think it ties to the race to the top, which is in my time at OpenAI, what I come to see as I'd come to appreciate the scaling hypothesis and as I'd come to appreciate kind of the importance of safety along with the scaling hypothesis. The first one I think OpenAI was getting on board with. The second one in a way had always been part of OpenAI's messaging. But over many years of the time that I spent there, I think I had a particular vision of how we should handle these things, how we should be brought out in the world, the kind of principles that the organization should have. And look, there were many, many discussions about should the company do this, should the company do that? There's a bunch of misinformation out there. People say we left because we didn't like the deal with Microsoft. False. Although, it was like a lot of discussion, a lot of questions about exactly how we do the deal with Microsoft. We left because we didn't like commercialization. That's not true. We built GPD-3, which was the model that was commercialized. I was involved in commercialization. It's more, again, about how do you do it? Civilization is going down this path to very powerful AI. What's the way to do it? That is cautious, straightforward, honest, that builds trust in the organization and in individuals. How do we get from here to there and how do we have a real vision for how to get it right? How can safety not just be something we say because it helps with recruiting. And I think at the end of the day, if you have a vision for that, forget about anyone else's vision. I don't want

to talk about anyone else's vision. If you have a vision for how to do it, you should go off and you should do that vision. It is incredibly unproductive to try and argue with someone else's vision. You might think they're not doing it the right way. You might think they're dishonest. Who knows? Maybe you're right, maybe you're not. But what you should do is you should take some people you trust and you should go off together and you should make your vision happen. And if your vision is compelling, if you can make it appeal to people, some combination of ethically in the market, if you can make a company that's a place people want to join, that engages in practices that people think are reasonable while managing to maintain its position in the ecosystem at the same time, if you do that, people will copy it. And the fact that you are doing it, especially the fact that you're doing it better than they are, causes them to change their behavior in a much more compelling way than if they're your boss and you're arguing with them. I don't know how to be any more specific about it than that, but I think it's generally very unproductive to try and get someone else's vision to look like your vision. It's much more productive to go off and do a clean experiment and say, "This is our vision, this is how we're going to do things. Your choice is you can ignore us, you can reject what we're doing or you can start to become more like us." And imitation is the sincerest form of flattery. And that plays out in the behavior of customers, that plays out in the behavior of the public, that plays out in the behavior of where people choose to work. And again, at the end, it's not about one company winning or another company winning. If we or another company are engaging in some practice that people find genuinely appealing, and I want it to be in substance, not just an appearance and I think researchers are sophisticated and they look at substance, and then other companies start copying that practice and they win because they copied that practice. That's great. That's success. That's like the race to the top. It doesn't matter who wins in the end as long as everyone is copying everyone else's good practices. One way I think of it is the thing we're all afraid of is the race to the bottom and the race to the bottom doesn't matter who wins because we all lose. In the most extreme world, we make this autonomous AI that the robots enslave us or whatever. That's half joking, but that is the most extreme thing that could happen. Then it doesn't matter which company was ahead. If instead you create a race to the top where people are competing to engage in good practices, then at the end of the day, it doesn't matter who ends up winning, it doesn't even matter who started the race to the top. The point isn't to be virtuous, the point is to get the system into a better equilibrium than it was before. And individual companies can play some role in doing this. Individual companies can help to start it, can help to accelerate it. And frankly, I think individuals at other companies have done this as well. The individuals that when we put out an RSP react by pushing harder to get something similar done at other companies, sometimes other companies do something that's we're like, "Oh, it's a good practice. We think that's good. We should adopt it too." The only difference is I think we try to be more forward leaning. We try and adopt more of these practices first and adopt them more quickly when others invent them. But I think this dynamic is what we should be pointing at and that I think it abstracts away the question of which company's winning, who trusts who. I think all these questions of drama are profoundly uninteresting and the thing that matters is the ecosystem that we all operate in and how to make that ecosystem better because that constrains all the players.

Lex Fridman
And so Anthropic is this kind of clean experiment built on a foundation of what concretely AI safety should look like?

Dario Amodei
Well, look, I'm sure we've made plenty of mistakes along the way. The perfect organization doesn't exist. It has to deal with the imperfection of a thousand employees. It has to deal with the imperfection of our leaders, including me. It has to deal with the imperfection of the people we've put to oversee the imperfection of the leaders like the board and the long-term benefit trust. It's all a set of imperfect people trying to aim imperfectly at some ideal that will never perfectly be achieved. That's what you sign up for. That's what it will always be. But imperfect doesn't mean you just give up. There's better and there's worse. And hopefully, we can do well enough that we can begin to build some practices that the whole industry engages in. And then my guess is that multiple of these companies will be successful. Anthropic will be successful. These other companies, like ones I've been at the past, will also be successful. And some will be more successful than others. That's less important than, again, that we align the incentives of the industry. And that happens partly through the race to the top, partly through things like RSP, partly through, again, selected surgical regulation.

Lex Fridman
You said talent density beats talent mass, so can you explain that? Can you expand on that?

Dario Amodei
Yeah.

Lex Fridman
Can you just talk about what it takes to build a great team of AI researchers and engineers?

Dario Amodei
This is one of these statements that's more true every month. Every month I see this statement as more true than I did the month before. So if I were to do a thought experiment, let's say you have a team of 100 people that are super smart, motivated and aligned with the mission and that's your company. Or you can have a team of a thousand people where 200 people are super smart, super aligned with the mission and then 800 people are, let's just say you pick 800 random big tech employees, which would you rather have? The talent mass is greater in the group of a thousand people. You have even a larger number of incredibly talented, incredibly aligned, incredibly smart people. But the issue is just that if every time someone super talented looks around, they see someone else super talented and super

dedicated, that sets the tone for everything. That sets the tone for everyone is super inspired to work at the same place. Everyone trusts everyone else. If you have a thousand or 10,000 people and things have really regressed, you are not able to do selection and you're choosing random people, what happens is then you need to put a lot of processes and a lot of guardrails in place just because people don't fully trust each other or you have to adjudicate political battles. There are so many things that slow down the org's ability to operate. And so we're nearly a thousand people and we've tried to make it so that as large a fraction of those thousand people as possible are super talented, super skilled, it's one of the reasons we've slowed down hiring a lot in the last few months. We grew from 300 to 800, I believe, I think in the first seven, eight months of the year and now we've slowed down. The last three months, we went from 800 to 900, 950, something like that. Don't quote me on the exact numbers, but I think there's an inflection point around a thousand and we want to be much more careful how we grow. Early on and now as well, we've hired a lot of physicists. Theoretical physicists can learn things really fast. Even more recently, as we've continued to hire that, we've really had a high bar on both the research side and the software engineering side, have hired a lot of senior people, including folks who used to be at other companies in this space, and we've just continued to be very selective. It's very easy to go from a hundred to a thousand, a thousand to 10,000 without paying attention to making sure everyone has a unified purpose. It's so powerful. If your company consists of a lot of different fiefdoms that all want to do their own thing, they're all optimizing for their own thing, it's very hard to get anything done. But if everyone sees the broader purpose of the company, if there's trust and there's dedication to doing the right thing, that is a superpower. That in itself I think can overcome almost every other disadvantage.

Lex Fridman
And to Steve Jobs, A players. A players want to look around and see other A players is another way of saying that.

Dario Amodei
Correct.

Lex Fridman
I don't know what that is about human nature, but it is demotivating to see people who are not obsessively driving towards a singular mission. And it is on the flip side of that, super motivating to see that. It's interesting. What's it take to be a great AI researcher or engineer from everything you've seen from working with so many amazing people?

Dario Amodei
Yeah. I think the number one quality, especially on the research side, but really both, is open mindedness. Sounds easy to be open-minded, right? You're just like, "Oh, I'm open to anything." But if I think about my own early history in this scaling hypothesis, I was seeing the same data others were seeing. I don't think I was a better programmer or better at

coming up with research ideas than any of the hundreds of people that I worked with. In some ways, I was worse. I've never precise programming of finding the bug, writing the GPU kernels. I could point you to a hundred people here who are better at that than I am. But the thing that I think I did have that was different was that I was just willing to look at something with new eyes. People said, "Oh, we don't have the right algorithms yet. We haven't come up with the right way to do things." And I was just like, "Oh, I don't know. This neural net has 30 million parameters. What if we gave it 50 million instead? Let's plot some graphs." That basic scientific mindset of like, "Oh man," I see some variable that I could change. What happens when it changes? Let's try these different things and create a graph. For even, this was the simplest thing in the world, change the number of, this wasn't PhD level experimental design, this was simple and stupid. Anyone could have done this if you just told them that it was important. It's also not hard to understand. You didn't need to be brilliant to come up with this. But you put the two things together and some tiny number of people, some single digit number of people have driven forward the whole field by realizing this. And it's often like that. If you look back at the discoveries in history, they're often like that. And so this open-mindedness and this willingness to see with new eyes that often comes from being newer to the field, often experience is a disadvantage for this, that is the most important thing. It's very hard to look for and test for, but I think it's the most important thing because when you find something, some really new way of thinking about things, when you have the initiative to do that, it's absolutely transformative.

Lex Fridman
And also be able to do kind of rapid experimentation and, in the face of that, be open-minded and curious and looking at the data with these fresh eyes and seeing what is it that it's actually saying. That applies in mechanistic interpretability.

Dario Amodei
It's another example of this. Some of the early work and mechanistic interpretability so simple, it's just no one thought to care about this question before.

Lex Fridman
You said what it takes to be a great AI researcher. Can we rewind the clock back, what advice would you give to people interested in AI? They're young, looking… What advice would you give to people interested in AI? They're young. Looking forward to how can I make an impact on the world?

Dario Amodei
I think my number one piece of advice is to just start playing with the models. Actually, I worry a little, this seems like obvious advice now. I think three years ago it wasn't obvious and people started by, "Oh, let me read the latest reinforcement learning paper." And you should do that as well, but now with wider availability of models and APIs, people are doing this more. But, I think just experiential knowledge. These models are new artifacts that no

one really understands and so getting experience playing with them. I would also say again, in line with the do something new, think in some new direction, there are all these things that haven't been explored. For example, mechanistic interpretability is still very new. It's probably better to work on that than it is to work on new model architectures, because it's more popular than it was before. There are probably 100 people working on it, but there aren't like 10,000 people working on it. And it's just this fertile area for study. There's so much low-hanging fruit, you can just walk by and you can pick things. For whatever reason, people aren't interested in it enough. I think there are some things around long horizon learning and long horizon tasks, where there's a lot to be done. I think evaluations, we're still very early in our ability to study evaluations, particularly for dynamic systems acting in the world. I think there's some stuff around multi-agent. Skate where the puck is going is my advice, and you don't have to be brilliant to think of it. All the things that are going to be exciting in five years, people even mention them as conventional wisdom, but it's just somehow there's this barrier that people don't double down as much as they could, or they're afraid to do something that's not the popular thing. I don't know why it happens, but getting over that barrier, that's my number one piece of advice.

Lex Fridman
Let's talk if we could a bit about post-training. So it seems that the modern post-training recipe has a little bit of everything. So supervised fine-tuning, RLHF, the constitutional AI with RLAIF-

Dario Amodei
Best acronym.

Lex Fridman
It's the, again, that naming thing. And then synthetic data. Seems like a lot of synthetic data, or at least trying to figure out ways to have high quality synthetic data. So if this is a secret sauce that makes Anthropic clause so incredible, how much of the magic is in the pre-training? How much of it is in the post-training?

Dario Amodei
Yeah. So first of all, we're not perfectly able to measure that ourselves. When you see some great character ability, sometimes it's hard to tell whether it came from pre-training or post-training. We developed ways to try and distinguish between those two, but they're not perfect. The second thing I would say is, when there is an advantage and I think we've been pretty good in general at RL, perhaps the best, although I don't know, I don't see what goes on inside other companies. Usually it isn't, "Oh my God, we have this secret magic method that others don't have." Usually it's like, "Well, we got better at the infrastructure so we could run it for longer," or, "We were able to get higher quality data," or, "We were able to filter our data better, or "We were able to combine these methods and practice." It's usually some boring matter of practice and trade craft. So when I think about how to do something

special in terms of how we train these models both, but even more so I really think of it a little more, again, as designing airplanes or cars. It's not just like, "Oh, man. I have the blueprint." Maybe that makes you make the next airplane. But there's some cultural trade craft of how we think about the design process that I think is more important than any particular gizmo we're able to invent.

Lex Fridman
Okay. Well, let me ask you about specific techniques. So first on RLHF, what do you think, just zooming out intuition, almost philosophy ... Why do you think RLHF works so well?

Dario Amodei
If I go back to the scaling hypothesis, one of the ways to skate the scaling hypothesis is, if you train for X and you throw enough compute at it, then you get X. And so RLHF is good at doing what humans want the model to do, or at least to state it more precisely doing what humans who look at the model for a brief period of time and consider different possible responses, what they prefer as the response, which is not perfect from both the safety and capabilities perspective, in that humans are often not able to perfectly identify what the model wants and what humans want in the moment may not be what they want in the long term. So there's a lot of subtlety there, but the models are good at producing what the humans in some shallow sense want. And it actually turns out that you don't even have to throw that much compute at it, because of another thing, which is this thing about a strong pre-trained model being halfway to anywhere. So once you have the pre-trained model, you have all the representations you need to get the model where you want it to go.

Lex Fridman
So do you think RLHF makes the model smarter, or just appear smarter to the humans?

Dario Amodei
I don't think it makes the model smarter. I don't think it just makes the model appear smarter. It's like RLHF bridges the gap between the human and the model. I could have something really smart that can't communicate at all. We all know people like this, people who are really smart but can't understand what they're saying. So I think RLHF just bridges that gap. I think it's not the only kind of RL we do. It's not the only kind of RL that will happen in the future. I think RL has the potential to make models smarter, to make them reason better, to make them operate better, to make them develop new skills even. And perhaps that could be done even in some cases with human feedback. But, the kind of RLHF we do today mostly doesn't do that yet, although we're very quickly starting to be able to.

Lex Fridman
But if you look at the metric of helpfulness, it increases that?

Dario Amodei
Yes. It also increases, what was this word in Leopold's essay, "unhobbling," where basically the models are hobbled and then you do various trainings to them to unhobble them. So I like that word, because it's a rare word. So I think RLHF unhobbles the models in some ways. And then there are other ways where that model hasn't yet been unhobbled and needs to unhobble.

Lex Fridman
If you can say in terms of cost, is pre-training the most expensive thing? Or is post-training creep up to that?

Dario Amodei
At the present moment, it is still the case that pre-training is the majority of the cost. I don't know what to expect in the future, but I could certainly anticipate a future where post-training is the majority of the cost.

Lex Fridman
In that future you anticipate, would it be the humans or the AI that's the costly thing for the post-training?

Dario Amodei
I don't think you can scale up humans enough to get high quality. Any kind of method that relies on humans and uses a large amount of compute, it's going to have to rely on some scaled supervision method, like debate or iterated amplification or something like that.

Lex Fridman
So on that super interesting set of ideas around constitutional AI, can you describe what it is as first detailed in December 2022 paper and beyond that. What is it?

Dario Amodei
Yes. So this was from two years ago. The basic idea is, so we describe what RLHF is. You have a model and you just sample from it twice. It spits out two possible responses, and you're like, "Human, which responses do you like better?" Or another variant of it is, "Rate this response on a scale of one to seven." So that's hard because you need to scale up human interaction and it's very implicit. I don't have a sense of what I want the model to do. I just have a sense of what this average of 1,000 humans wants the model to do. So two ideas. One is, could the AI system itself decide which response is better? Could you show the AI system these two responses and ask which response is better? And then second, well, what criterion should the AI use? And so then there's this idea, you have a single document, a constitution if you will, that says, these are the principles the model should be using to respond. And the AI system reads those reads principles as well as reading the environment and the response. And it says, "Well, how good did the AI model do?" It's

basically a form of self-play. You're training the model against itself. And so the AI gives the response and then you feed that back into what's called the preference model, which in turn feeds the model to make it better. So you have this triangle of the AI, the preference model, and the improvement of the AI itself.

Lex Fridman
And we should say that in the constitution, the set of principles are human interpretable. They're-

Dario Amodei
Yeah. Yeah. It's something both the human and the AI system can read. So it has this nice translatability or symmetry. In practice, we both use a model constitution and we use RLHF and we use some of these other methods. So it's turned into one tool in a toolkit, that both reduces the need for RLHF and increases the value we get from using each data point of RLHF. It also interacts in interesting ways with future reasoning type RL methods. So it's one tool in the toolkit, but I think it is a very important tool.

Lex Fridman
Well, it's a compelling one to us humans. Thinking about the founding fathers and the founding of the United States. The natural question is who and how do you think it gets to define the constitution, the set of principles in the constitution?

Dario Amodei
Yeah. So I'll give a practical answer and a more abstract answer. I think the practical answer is look in practice, models get used by all kinds of different customers. And so you can have this idea where the model can have specialized rules or principles. We fine tune versions of models implicitly. We've talked about doing it explicitly having special principles that people can build into the models. So from a practical perspective, the answer can be very different from different people. A customer service agent behaves very differently from a lawyer and obeys different principles. But, I think at the base of it, there are specific principles that models have to obey. I think a lot of them are things that people would agree with. Everyone agrees that we don't want models to present these CBRN risks. I think we can go a little further and agree with some basic principles of democracy and the rule of law. Beyond that, it gets very uncertain and there our goal is generally for the models to be more neutral, to not espouse a particular point of view and more just be wise agents or advisors that will help you think things through and will present possible considerations. But don't express strong or specific opinions.

Lex Fridman
OpenAI released a model spec where it clearly, concretely defines some of the goals of the model and specific examples like AB, how the model should behave. Do you find that interesting? By the way I should mention, I believe the brilliant John Schulman was a part of

that. He’s now at Anthropic. Do you think this is a useful direction? Might Anthropic release a model spec as well?

Dario Amodei
Yeah. So I think that’s a pretty useful direction. Again, it has a lot in common with constitutional AI. So again, another example of a race to the top. We have something that we think a better and more responsible way of doing things. It’s also a competitive advantage. Then others discover that it has advantages and then start to do that thing. We then no longer have the competitive advantage, but it’s good from the perspective that now everyone has adopted a positive practice that others were not adopting. And so our response to that is, “Well, looks like we need a new competitive advantage in order to keep driving this race upwards.” So that’s how I generally feel about that. I also think every implementation of these things is different. So there were some things in the model spec that were not in constitutional AI, and so we can always adopt those things or at least learn from them. So again, I think this is an example of the positive dynamic that I think we should all want the field to have.

Lex Fridman
Let’s talk about the incredible essay Machines of Loving Grace. I recommend everybody read it. It’s a long one.

Dario Amodei
It is rather long.

Lex Fridman
Yeah. It’s really refreshing to read concrete ideas about what a positive future looks like. And you took a bold stance because it’s very possible that you might be wrong on the dates or the specific applications-

Dario Amodei
Oh, yeah. I’m fully expecting to well, definitely be wrong about all the details. I might be just spectacularly wrong about the whole thing and people will laugh at me for years. That’s just how the future works.

Lex Fridman
So you provided a bunch of concrete positive impacts of AI and how exactly a super intelligent AI might accelerate the rate of breakthroughs in, for example, biology and chemistry, that would then lead to things like we cure most cancers, prevent all infectious disease, double the human lifespan and so on. So let’s talk about this essay first. Can you give a high-level vision of this essay? And what are the key takeaways that people have?

Dario Amodei
Yeah. I have spent a lot of time, and in Anthropic has spent a lot of effort on how do we address the risks of AI? How do we think about those risks? We're trying to do a race to the top, what that requires us to build all these capabilities and the capabilities are cool. But, a big part of what we're trying to do is address the risks. And the justification for that is like, well, all these positive things, the market is this very healthy organism. It's going to produce all the positive things. The risks? I don't know, we might mitigate them, we might not. And so we can have more impact by trying to mitigate the risks. But, I noticed that one flaw in that way of thinking, and it's not a change in how seriously I take the risks. It's maybe a change in how I talk about them, is that no matter how logical or rational, that line of reasoning that I just gave might be. If you only talk about risks, your brain only thinks about risks. And so, I think it's actually very important to understand, what if things do go well? And the whole reason we're trying to prevent these risks is not because we're afraid of technology, not because we want to slow it down. It's because if we can get to the other side of these risks, if we can run the gauntlet successfully, to put it in stark terms, then on the other side of the gauntlet are all these great things. And these things are worth fighting for. And these things can really inspire people. And I think I imagine, because … Look, you have all these investors, all these VCs, all these AI companies talking about all the positive benefits of AI. But as you point out, it's weird. There's actually a dearth of really getting specific about it. There's a lot of random people on Twitter posting these gleaming cities and this just vibe of grind, accelerate harder, kick out the … It's just this very aggressive ideological. But then you're like, "Well, what are you actually excited about?" And so, I figured that I think it would be interesting and valuable for someone who's actually coming from the risk side to try and really make a try at explaining what the benefits are, both because I think it's something we can all get behind and I want people to understand. I want them to really understand that this isn't Doomers versus Accelerationists. This is that, if you have a true understanding of where things are going with AI, and maybe that's the more important axis, AI is moving fast versus AI is not moving fast, then you really appreciate the benefits and you really want humanity or civilization to seize those benefits. But, you also get very serious about anything that could derail them.

Lex Fridman
So I think the starting point is to talk about what this Powerful AI, which is the term you like to use, most of the world uses AGI, but you don't like the term, because it's basically has too much baggage, it's become meaningless. It's like we're stuck with the terms whether we like them or not.

Dario Amodei
Maybe we're stuck with the terms and my efforts to change them are futile.

Lex Fridman
It's admirable.

Dario Amodei
I'll tell you what else I don't … This is a pointless semantic point, but I keep talking about it-

Lex Fridman
It's back to naming again.

Dario Amodei
I'm just going to do it once more. I think it's a little like, let's say it was like 1995 and Moore's law is making the computers faster. And for some reason there had been this verbal tick that everyone was like, "Well, someday we're going to have supercomputers. And supercomputers are going to be able to do all these things that … Once we have supercomputers, we'll be able to sequence the genome, we'll be able to do other things." And so. One, it's true, the computers are getting faster and as they get faster, they're going to be able to do all these great things. But there's, there's no discrete point at which you had a supercomputer and previous computers were no. "Supercomputer" is a term we use, but it's a vague term to just describe computers that are faster than what we have today. There's no point at which you pass the threshold and you're like, "Oh, my God! We're doing a totally new type of computation and new … And so I feel that way about AGI. There's just a smooth exponential. And if by AGI you mean AI is getting better and better, and gradually it's going to do more and more of what humans do until it's going to be smarter than humans, and then it's going to get smarter even from there, then yes, I believe in AGI. But, if AGI is some discrete or separate thing, which is the way people often talk about it, then it's a meaningless buzzword.

Lex Fridman
To me, it's just a platonic form of a powerful AI, exactly how you define it. You define it very nicely, so on the intelligence axis, it's just on pure intelligence, it's smarter than a Nobel Prize winner as you describe across most relevant disciplines. So okay, that's just intelligence. So it's both in creativity and be able to generate new ideas, all that kind of stuff in every discipline, Nobel Prize winner in their prime. It can use every modality, so this is self-explanatory, but just operate across all the modalities of the world. It can go off for many hours, days and weeks to do tasks and do its own detailed planning and only ask you help when it's needed. This is actually interesting. I think in the essay you said … Again, it's a bet that it's not going to be embodied, but it can control embodied tools. So it can control tools, robots, laboratory equipment., the resource used to train it can then be repurposed to run millions of copies of it, and each of those copies would be independent that could do their own independent work. So you can do the cloning of the intelligence systems.

Dario Amodei
Yeah. Yeah. You might imagine from outside the field that there's only one of these, right? You've only made one. But the truth is that the scale up is very quick. We do this today,. We make a model, and then we deploy thousands, maybe tens of thousands of instances of it. I

think by the time, certainly within two to three years, whether we have these super powerful AIs or not, clusters are going to get to the size where you'll be able to deploy millions of these. And they'll be faster than humans. And so, if your picture is, "Oh, we'll have one and it'll take a while to make them," my point there was, no. Actually you have millions of them right away.

Lex Fridman
And in general they can learn and act 10 to 100 times faster than humans. So that's a really nice definition of powerful AI. Okay, so that. But, you also write that, "Clearly such an entity would be capable of solving very difficult problems very fast, but it is not trivial to figure out how fast. Two "extreme" positions both seem false to me." So the singularity is on the one extreme and the opposite and the other extreme. Can you describe each of the extremes?

Dario Amodei
Yeah.

Lex Fridman
So why?

Dario Amodei
So yeah. Let's describe the extreme. So one extreme would be, "Well, look. If we look at evolutionary history like there was this big acceleration, where for hundreds of thousands of years we just had single-celled organisms, and then we had mammals, and then we had apes. And then that quickly turned to humans. Humans quickly built industrial civilization." And so, this is going to keep speeding up and there's no ceiling at the human level. Once models get much, much smarter than humans, they'll get really good at building the next models. And if you write down a simple differential equation, like this is an exponential … And so what's going to happen is that models will build faster models. Models will build faster models. And those models will build nanobots that can take over the world and produce much more energy than you could produce otherwise. And so, if you just kind of solve this abstract differential equation, then like five days after we build the first AI that's more powerful than humans, then the world will be filled with these AIs in every possible technology that could be invented, like will be invented. I'm caricaturing this a little bit, but I think that's one extreme. And the reason that I think that's not the case is that, one, I think they just neglect the laws of physics. It's only possible to do things so fast in the physical world. Some of those loops go through producing faster hardware. It takes a long time to produce faster hardware. Things take a long time. There's this issue of complexity. I think no matter how smart you are, people talk about, "Oh, we can make models of biological systems that'll do everything the biological systems … " Look, I think computational modeling can do a lot. I did a lot of computational modeling when I worked in biology. But just there are a lot of things that you can't predict how … They're complex enough that just

iterating, just running the experiment is going to beat any modeling, no matter how smart the system doing the modeling is.

Lex Fridman
Well, even if it's not interacting with the physical world, just the modeling is going to be hard?

Dario Amodei
Yeah. Well, the modeling is going to be hard and getting the model to match the physical world is going to be

Lex Fridman
All right. So it does have to interact with the physical world to verify.

Dario Amodei
But you just look at even the simplest problems. I think I talk about The Three-Body Problem or simple chaotic prediction, or predicting the economy. It's really hard to predict the economy two years out. Maybe the case is humans can predict what's going to happen in the economy next quarter, or they can't really do that. Maybe a AI that's a zillion times smarter can only predict it out a year or something, instead of … You have these exponential increase in computer intelligence for linear increase in ability to predict. Same with again, like biological molecules interacting. You don't know what's going to happen when you perturb a complex system. You can find simple parts in it, if you're smarter, you're better at finding these simple parts. And then I think human institutions, human institutions are really difficult. It's been a hard to get people. I won't give specific examples, but it's been hard to get people to adopt even the technologies that we've developed, even ones where the case for their efficacy is very, very strong. People have concerns. They think things are conspiracy theories. It's just been very difficult. It's also been very difficult to get very simple things through the regulatory system. And I don't want to disparage anyone who works in regulatory systems of any technology. There are hard they have to deal with. They have to save lives. But the system as a whole, I think makes some obvious trade-offs that are very far from maximizing human welfare. And so, if we bring AI systems into these human systems, often the level of intelligence may just not be the limiting factor. It just may be that it takes a long time to do something. Now, if the AI system circumvented all governments, if it just said, "I'm dictator of the world and I'm going to do whatever," some of these things it could do. Again, the things have to do with complexity. I still think a lot of things would take a while. I don't think it helps that the AI systems can produce a lot of energy or go to the moon. Like some people in comments responded to the essay saying the AI system can produce a lot of energy and smarter AI systems. That's missing the point. That kind of cycle doesn't solve the key problems that I'm talking about here. So I think a bunch of people missed the point there. But even if it were completely unaligned and could get around all these human obstacles it would have trouble. But again, if you want this to be

an AI system that doesn't take over the world, that doesn't destroy humanity, then basically it's going to need to follow basic human laws. If we want to have an actually good world, we're going to have to have an AI system that interacts with humans, not one that creates its own legal system, or disregards all the laws or all of that. So as inefficient as these processes are, we're going to have to deal with them, because there needs to be some popular and democratic legitimacy in how these systems are rolled out. We can't have a small group of people who are developing these systems say, "This is what's best for everyone." I think it's wrong, and I think in practice it's not going to work anyway. So you put all those things together and we're not going change the world and upload everyone in five minutes. A, I don't think it's going to happen and B, to the extent that it could happen.,It's not the way to lead to a good world. So that's on one side. On the other side, there's another set of perspectives, which I have actually in some ways more sympathy for, which is, look, we've seen big productivity increases before. Economists are familiar with studying the productivity increases that came from the computer revolution and internet revolution. And generally those productivity increases were underwhelming. They were less than you might imagine. There was a quote from Robert Solow, "You see the computer revolution everywhere except the productivity statistics." So why is this the case? People point to the structure of firms, the structure of enterprises, how slow it's been to roll out our existing technology to very poor parts of the world, which I talk about in the essay. How do we get these technologies to the poorest parts of the world that are behind on cell phone technology, computers, medicine, let alone newfangled AI that hasn't been invented yet. So you could have a perspective that's like, "Well, this is amazing technically, but it's all or nothing burger. I think Tyler Cowen who wrote something in response to my essay has that perspective. I think he thinks the radical change will happen eventually, but he thinks it'll take 50 or 100 years. And you could have even more static perspectives on the whole thing. I think there's some truth to it. I think the time scale is just too long and I can see it. I can actually see both sides with today's AI. So a lot of our customers are large enterprises who are used to doing things a certain way. I've also seen it in talking to governments, right? Those are prototypical institutions, entities that are slow to change. But, the dynamic I see over and over again is yes, it takes a long time to move the ship. Yes. There's a lot of resistance and lack of understanding. But, the thing that makes me feel that progress will in the end happen moderately fast, not incredibly fast, but moderately fast, is that you talk to … What I find is I find over and over again, again in large companies, even in governments which have been actually surprisingly forward leaning, you find two things that move things forward. One, you find a small fraction of people within a company, within a government, who really see the big picture, who see the whole scaling hypothesis, who understand where AI is going, or at least understand where it's going within their industry. And there are a few people like that within the current US government who really see the whole picture. And those people see that this is the most important thing in the world until they agitate for it. And the thing they alone are not enough to succeed, because there are a small set of people within a large organization. But, as the technology starts to roll out, as it succeeds in some places in the folks who are most willing to adopt it, the specter of competition gives them a wind at their backs, because they can point within their large organization. They can say, “Look, these other guys are doing this.” One bank can say, “Look, this newfangled hedge fund is doing this thing. They’re going to eat our lunch.” In the US, we can say we’re afraid China’s going to get there before we are. And that combination, the specter of competition plus a few visionaries within these, the organizations that in many ways are sclerotic, you put those two things together and it actually makes something happen. It’s interesting. It’s a balanced fight between the two, because inertia is very powerful, but eventually over enough time, the innovative approach breaks through. And I’ve seen that happen. I’ve seen the arc of that over and over again, and it’s like the barriers are there, the barriers to progress, the complexity, not knowing how to use the model, how to deploy them are there. And for a bit it seems like they’re going to last forever, change doesn’t happen. But, then eventually change happens and always comes from a few people. I felt the same way when I was an advocate of the scaling hypothesis within the AI field itself and others didn’t get it. It felt like no one would ever get it. Then it felt like we had a secret almost no one ever had. And then, a couple years later, everyone has the secret. And so, I think that’s how it’s going to go with deployment AI in the world. The barriers are going to fall apart gradually and then all at once. And so, I think this is going to be more, and this is just an instinct. I could easily see how I’m wrong. I think it’s going to be more five or 10 years, as I say in the essay than it’s going to be 50 or 100 years. I also think it’s going to be five or 10 years more than it’s going to be five or 10 hours, because I’ve just seen how human systems work. And I think a lot of these people who write down these differential equations, who say AI is going to make more powerful AI, who can’t understand how it could possibly be the case that these things won’t change so fast. I think they don’t understand these things.

Lex Fridman
So what to you is the timeline to where we achieve AGI - aka powerful AI - aka super useful AI?

Dario Amodei
I’m going to start calling it that.

Lex Fridman
It’s a debate about naming. On pure intelligence smarter than a Nobel Prize winner in every relevant discipline and all the things we’ve said. Modality, can go and do stuff on its own for days, weeks, and do biology experiments on its own in one … You know what? Let’s just stick to biology, because you sold me on the whole biology and health section. And that’s so exciting from just … I was getting giddy from a scientific perspective. It made me want to be a biologist.

Dario Amodei
So no,. No. This was the feeling I had when I was writing it, that it’s like, this would be such a beautiful future if we can just make it happen. If we can just get the landmines out of the

way and make it happen. There's so much beauty and elegance and moral force behind it if we can just … And it's something we should all be able to agree on. As much as we fight about all these political questions, is this something that could actually bring us together? But you were asking when will we get this?

Lex Fridman
When? When do you think? Just putting numbers on the table.

Dario Amodei
This is, of course, the thing I've been grappling with for many years, and I'm not at all confident. If I say 2026 or 2027, there will be a zillion people on Twitter who will be like, "AI CEO said 2026, 2020 … " and it'll be repeated for the next two years that this is definitely when I think it's going to happen. So whoever's exerting these clips will crop out the thing I just said and only say the thing I'm about to say. But I'll just say it anyway-

Lex Fridman
Have fun with it.

Dario Amodei
So if you extrapolate the curves that we've had so far. Right? If you say, "Well, I don't know. We're starting to get to PhD level, and last year we were at undergraduate level and the year before we were at the level of a high school student." Again, you can quibble with at what tasks and for what we're still missing modalities, but those are being added. Computer use was added, like ImageEn was added, image generation has been added. And this is totally unscientific, but if you just eyeball the rate at which these capabilities are increasing, it does make you think that we'll get there by 2026 or 2027. Again, lots of things could derail it. We could run out of data. We might not be able to scale clusters as much as we want. Maybe Taiwan gets blown up or something, and then we can't produce as many GPUs as we want. So there are all- Then we can't produce as many GPUs as we want. So there are all kinds of things that could derail the whole process. So I don't fully believe the straight line extrapolation, but if you believe the straight line extrapolation, we'll get there in 2026 or 2027. I think the most likely is that there are some mild delay relative to that. I don't know what that delay is, but I think it could happen on schedule. I think there could be a mild delay. I think there are still worlds where it doesn't happen in a hundred years. The number of those worlds is rapidly decreasing. We are rapidly running out of truly convincing blockers, truly compelling reasons why this will not happen in the next few years. There were a lot more in 2020, although my guess, my hunch at that time was that we'll make it through all those blockers. So sitting as someone who has seen most of the blockers cleared out of the way, I suspect, my hunch, my suspicion is that the rest of them will not block us. But look, at the end of the day, I don't want to represent this as a scientific prediction. People call them scaling laws. That's a misnomer. Like Moore's law is a

misnomer. Moore's laws, scaling laws, they're not laws of the universe. They're empirical regularities. I am going to bet in favor of them continuing, but I'm not certain of that.

Lex Fridman
So you extensively described sort of the compressed 21st century, how AGI will help set forth a chain of breakthroughs in biology and medicine that help us in all these kinds of ways that I mentioned. What are the early steps it might do? And by the way, I asked Claude good questions to ask you and Claude told me to ask, what do you think is a typical day for a biologist working on AGI look like in this future?

Dario Amodei
Yeah, yeah.

Lex Fridman
Claude is curious.

Dario Amodei
Well, let me start with your first questions and then I'll answer that. Claude wants to know what's in his future, right?

Lex Fridman
Exactly.

Dario Amodei
Who am I going to be working with?

Lex Fridman
Exactly.

Dario Amodei
So I think one of the things when I went hard on in the essay is let me go back to this idea of, because it's really had an impact on me, this idea that within large organizations and systems, there end up being a few people or a few new ideas who cause things to go in a different direction than they would've before who kind of disproportionately affect the trajectory. There's a bunch of the same thing going on, right? If you think about the health world, there's like trillions of dollars to pay out Medicare and other health insurance and then the NIH is 100 billion. And then if I think of the few things that have really revolutionized anything, it could be encapsulated in a small fraction of that. And so when I think of where will AI have an impact, I'm like, "Can AI turn that small fraction into a much larger fraction and raise its quality?" And within biology, my experience within biology is that the biggest problem of biology is that you can't see what's going on. You have very little ability to see what's going on and even less ability to change it, right? What you have is this. From this,

you have to infer that there's a bunch of cells that within each cell is 3 billion base pairs of DNA built according to a genetic code. And there are all these processes that are just going on without any ability of us on unaugmented humans to affect it. These cells are dividing. Most of the time that's healthy, but sometimes that process goes wrong and that's cancer. The cells are aging, your skin may change color, develops wrinkles as you age, and all of this is determined by these processes. All these proteins being produced, transported to various parts of the cells binding to each other. And in our initial state about biology, we didn't even know that these cells existed. We had to invent microscopes to observe the cells. We had to invent more powerful microscopes to see below the level of the cell to the level of molecules. We had to invent X-ray crystallography to see the DNA. We had to invent gene sequencing to read the DNA. Now we had to invent protein folding technology to predict how it would fold and how these things bind to each other. We had to invent various techniques for now we can edit the DNA as of with CRISPR as of the last 12 years. So the whole history of biology, a whole big part of the history is basically our ability to read and understand what's going on and our ability to reach in and selectively change things. And my view is that there's so much more we can still do there. You can do CRISPR, but you can do it for your whole body. Let's say I want to do it for one particular type of cell and I want the rate of targeting the wrong cell to be very low. That's still a challenge. That's still things people are working on. That's what we might need for gene therapy for certain diseases. The reason I'm saying all of this, it goes beyond this to gene sequencing, to new types of nanomaterials for observing what's going on inside cells, for antibody drug conjugates. The reason I'm saying all this is that this could be a leverage point for the AI systems, right? That the number of such inventions, it's in the mid double digits or something, mid double digits, maybe low triple digits over the history of biology. Let's say I have a million of these AIs like can they discover a thousand working together or can they discover thousands of these very quickly and does that provide a huge lever? Instead of trying to leverage two trillion a year we spend on Medicare or whatever, can we leverage the 1 billion a year that's spent to discover but with much higher quality? And so what is it like being a scientist that works with an AI system? The way I think about it actually is, well, so I think in the early stages, the AIs are going to be like grad students. You're going to give them a project. You're going to say, "I'm the experienced biologist. I've set up the lab." The biology professor or even the grad students themselves will say, "Here's what you can do with an AI… AI system, I'd like to study this." And the AI system, it has all the tools. It can look up all the literature to decide what to do. It can look at all the equipment. It can go to a website and say, "Hey, I'm going to go to Thermo Fisher or whatever the dominant lab equipment company is today. My time was Thermo Fisher. I'm going to order this new equipment to do this. I'm going to run my experiments. I'm going to write up a report about my experiments. I'm going to inspect the images for contamination. I'm going to decide what the next experiment is. I'm going to write some code and run a statistical analysis. All the things a grad student would do that'll be a computer with an AI that the professor talks to every once in a while and it says, "This is what you're going to do today." The AI system comes to it with questions. When it's necessary to run the lab equipment, it may be limited in some ways. It may have to hire a human lab assistant to do the experiment and explain how to do it or it could use advances in lab automation that are gradually being developed or have been developed over the last decade or so and will continue to be developed. And so it'll look like there's a human professor and 1,000 AI grad students and if you go to one of these Nobel Prize winning biologists or so, you'll say, "Okay, well, you had like 50 grad students. Well, now you have 1,000 and they're smarter than you are by the way." Then I think at some point it'll flip around where the AI systems will be the PIs, will be the leaders, and they'll be ordering humans or other AI systems around. So I think that's how it'll work on the research side.

Lex Fridman
And there would be the inventors of a CRISPR type technology.

Dario Amodei
They would be the inventors of a CRISPR type technology. And then I think, as I say in the essay, we'll want to turn, probably turning loose is the wrong term, but we'll want to harness the AI systems to improve the clinical trial system as well. There's some amount of this that's regulatory, that's a matter of societal decisions and that'll be harder. But can we get better at predicting the results of clinical trials? Can we get better at statistical design so that clinical trials that used to require 5,000 people and therefore needed $100 million in a year to enroll them, now they need 500 people in two months to enroll them? That's where we should start. And can we increase the success rate of clinical trials by doing things in animal trials that we used to do in clinical trials and doing things in simulations that we used to do in animal trials? Again, we won't be able to simulate at all. AI is not God, but can we shift the curve substantially and radically? So I don't know, that would be my picture.

Lex Fridman
Doing in vitro and doing it. I mean you're still slowed down. It still takes time, but you can do it much, much faster.

Dario Amodei
Yeah, yeah. Can we just one step at a time and can that add up to a lot of steps? Even though though we still need clinical trials, even though we still need laws, even though the FDA and other organizations will still not be perfect, can we just move everything in a positive direction and when you add up all those positive directions, do you get everything that was going to happen from here to 2100 instead happens from 2027 to 2032 or something?

Lex Fridman
Another way that I think the world might be changing with AI even today, but moving towards this future of the powerful super useful AI is programming. So how do you see the nature of programming because it's so intimate to the actual act of building AI. How do you see that changing for us humans?

Dario Amodei
I think that's going to be one of the areas that changes fastest for two reasons. One, programming is a skill that's very close to the actual building of the AI. So the farther a skill is from the people who are building the AI, the longer it's going to take to get disrupted by the AI. I truly believe that AI will disrupt agriculture. Maybe it already has in some ways, but that's just very distant from the folks who are building AI, and so I think it's going to take longer. But programming is the bread and butter of a large fraction of the employees who work at Anthropic and at the other companies, and so it's going to happen fast. The other reason it's going to happen fast is with programming, you close the loop both when you're training the model and when you're applying the model. The idea that the model can write the code means that the model can then run the code and then see the results and interpret it back. And so it really has an ability unlike hardware, unlike biology, which we just discussed, the model has an ability to close the loop. And so I think those two things are going to lead to the model getting good at programming very fast. As I saw on typical real-world programming tasks, models have gone from 3% in January of this year to 50% in October of this year. So we're on that S-curve where it's going to start slowing down soon because you can only get to 100%. But I would guess that in another 10 months, we'll probably get pretty close. We'll be at least 90%. So again, I would guess, I don't know how long it'll take, but I would guess again, 2026, 2027 Twitter people who crop out these numbers and get rid of the caveats, I don't know. I don't like you, go away. I would guess that the kind of task that the vast majority of coders do, AI can probably, if we make the task very narrow, just write code, AI systems will be able to do that. Now that said, I think comparative advantage is powerful. We'll find that when AIs can do 80% of a coder's job, including most of it that's literally write code with a given spec, we'll find that the remaining parts of the job become more leveraged for humans, right? Humans, there'll be more about high level system design or looking at the app and is it architected well and the design and UX aspects and eventually AI will be able to do those as well. That's my vision of the powerful AI system. But I think for much longer than we might expect, we will see that small parts of the job that humans still do will expand to fill their entire job in order for the overall productivity to go up. That's something we've seen. It used to be that writing and editing letters was very difficult and writing the print was difficult. Well, as soon as you had word processors and then computers and it became easy to produce work and easy to share it, then that became instant and all the focus was on the ideas. So this logic of comparative advantage that expands tiny parts of the tasks to large parts of the tasks and creates new tasks in order to expand productivity, I think that's going to be the case. Again, someday AI will be better at everything and that logic won't apply, and then humanity will have to think about how to collectively deal with that and we're thinking about that every day and that's another one of the grand problems to deal with aside from misuse and autonomy and we should take it very seriously. But I think in the near term, and maybe even in the medium term, medium term like 2, 3, 4 years, I expect that humans will continue to have a huge role and the nature of programming will change, but programming as a role, programming as a job will not change. It'll just be less writing things line by line and it'll be more macroscopic.

Lex Fridman
And I wonder what the future of IDEs looks like. So the tooling of interacting with AI systems, this is true for programming and also probably true for in other contexts like computer use, but maybe domain specific, like we mentioned biology, it probably needs its own tooling about how to be effective. And then programming needs its own tooling. Is Anthropic going to play in that space of also tooling potentially?

Dario Amodei
I'm absolutely convinced that powerful IDEs, that there's so much low-hanging fruit to be grabbed there that right now it's just like you talk to the model and it talks back. But look, I mean IDEs are great at lots of static analysis of so much is possible with static analysis like many bugs you can find without even writing the code. Then IDEs are good for running particular things, organizing your code, measuring coverage of unit tests. There's so much that's been possible with a normal IDEs. Now you add something like, well, the model can now write code and run code. I am absolutely convinced that over the next year or two, even if the quality of the models didn't improve, that there would be enormous opportunity to enhance people's productivity by catching a bunch of mistakes, doing a bunch of grunt work for people, and that we haven't even scratched the surface. Anthropic itself, I mean you can't say no… It's hard to say what will happen in the future. Currently, we're not trying to make such IDEs ourself, rather we're powering the companies like Cursor or Kognition or some of the other expo in the security space, others that I could mention as well that are building such things themselves on top of our API and our view has been let 1,000 flowers bloom. We don't internally have the resources to try all these different things. Let's let our customers try it and we will see who succeeds and maybe different customers will succeed in different ways. So I both think this is super promising and Anthropic isn't eager to, at least right now, compete with all our companies in this space and maybe never.

Lex Fridman
Yeah, it's been interesting to watch Cursor try to integrate cloud successfully because it's actually fascinating how many places it can help the programming experience. It's not as trivial.

Dario Amodei
It is really astounding. I feel like as a CEO, I don't get to program that much, and I feel like if six months from now I go back, it'll be completely unrecognizable to me.

Lex Fridman
Exactly. In this world with super powerful AI that's increasingly automated, what's the source of meaning for us humans? Work is a source of deep meaning for many of us. Where do we find the meaning?

Dario Amodei
This is something that I've written about a little bit in the essay, although I actually give it a bit short shrift, not for any principled reason, but this essay, if you believe it was originally going to be two or three pages, I was going to talk about it at all hands. And the reason I realized it was an important underexplored topic is that I just kept writing things and I was just like, "Oh man, I can't do this justice." And so the thing ballooned to 40 or 50 pages and then when I got to the work and meaning section, I'm like, "Oh man, this isn't going to be 100 pages." I'm going to have to write a whole other essay about that. But meaning is actually interesting because you think about the life that someone lives or something, or let's say you were to put me in, I don't know, like a simulated environment or something where I have a job and I'm trying to accomplish things and I don't know, I do that for 60 years and then you're like, "Oh, oops, this was actually all a game," right? Does that really kind of rob you of the meaning of the whole thing? I still made important choices, including moral choices. I still sacrificed. I still had to gain all these skills or just a similar exercise. Think back to one of the historical figures who discovered electromagnetism or relativity or something. If you told them, "Well, actually 20,000 years ago, some alien on this planet discovered this before you did," does that rob the meaning of the discovery? It doesn't really seem like it to me, right? It seems like the process is what matters and how it shows who you are as a person along the way and how you relate to other people and the decisions that you make along the way. Those are consequential. I could imagine if we handle things badly in an AI world, we could set things up where people don't have any long-term source of meaning or any, but that's more a set of choices we make that's more a set of the architecture of society with these powerful models. If we design it badly and for shallow things, then that might happen. I would also say that most people's lives today, while admirably, they work very hard to find meaning in those lives. Like look, we who are privileged and who are developed these technologies, we should have empathy for people not just here, but in the rest of the world who spend a lot of their time scraping by to survive, assuming we can distribute the benefits of this technology to everywhere, their lives are going to get a hell of a lot better and meaning will be important to them as it is important to them now. But we should not forget the importance of that and that the idea of meaning as the only important thing is in some ways an artifact of a small subset of people who have been economically fortunate. But I think all of that said, I think a world is possible with powerful AI that not only has as much meaning for everyone, but that has more meaning for everyone that can allow everyone to see worlds and experiences that it was either possible for no one to see or a possible for very few people to experience. So I am optimistic about meaning. I worry about economics and the concentration of power. That's actually what I worry about more. I worry about how do we make sure that that fair world reaches everyone. When things have gone wrong for humans, they've often gone wrong because humans mistreat other humans. That is maybe in some ways even more than the autonomous risk of AI or the question of meaning. That is the thing I worry about most, the concentration of power, the abuse of power, structures like autocracies and dictatorships where a small number of people exploits a large number of people. I'm very worried about that.

Lex Fridman
And AI increases the amount of power in the world, and if you concentrate that power and abuse that power, it can do immeasurable damage.

Dario Amodei
Yes, it's very frightening. It's very frightening.

Lex Fridman
Well, I encourage highly encourage people to read the full essay. That should probably be a book or a sequence of essays because it does paint a very specific future. And I could tell the later sections got shorter and shorter because you started to probably realize that this is going to be a very long essay if you keep going.

Dario Amodei
One, I realized it would be very long, and two, I'm very aware of and very much tried to avoid just being, I don't know what the term for it is, but one of these people who's overconfident and has an opinion on everything and says a bunch of stuff and isn't an expert, I very much tried to avoid that. But I have to admit, once I got to biology sections, I wasn't an expert. And so as much as I expressed uncertainty, probably I said a bunch of things that were embarrassing or wrong.

Lex Fridman
Well, I was excited for the future you painted, and thank you so much for working hard to build that future and thank you for talking to me, Dario.

Dario Amodei
Thanks for having me. I just hope we can get it right and make it real. And if there's one message I want to send, it's that to get all this stuff right, to make it real, we both need to build the technology, build the companies, the economy around using this technology positively, but we also need to address the risks because those risks are in our way. They're landmines on the way from here to there, and we have to diffuse those landmines if we want to get there.

Lex Fridman
It's a balance like all things in life.

Dario Amodei
Like all things.

Lex Fridman
Thank you. Thanks for listening to this conversation with Dario Amodei. And now, dear friends, here's Amanda Askell. You are a philosopher by training. So what sort of questions

did you find fascinating through your journey in philosophy in Oxford and NYU and then switching over to the AI problems at OpenAI and Anthropic?

Lex Fridman

I think philosophy is actually a really good subject if you are fascinated with everything because there's a philosophy all of everything. So if you do philosophy of mathematics for a while and then you decide that you're actually really interested in chemistry, you can do philosophy of chemistry for a while, you can move into ethics or philosophy of politics. I think towards the end, I was really interested in ethics primarily. So that was what my PhD was on. It was on a kind of technical area of ethics, which was ethics where worlds contain infinitely many people, strangely, a little bit less practical on the end of ethics. And then I think that one of the tricky things with doing a PhD in ethics is that you're thinking a lot about the world, how it could be better, problems, and you're doing a PhD in philosophy. And I think when I was doing my PhD, I was like this is really interesting. It's probably one of the most fascinating questions I've ever encountered in philosophy and I love it, but I would rather see if I can have an impact on the world and see if I can do good things. And I think that was around the time that AI was still probably not as widely recognized as it is now. That was around 2017, 2018. It had been following progress and it seemed like it was becoming kind of a big deal. And I was basically just happy to get involved and see if I could help because I was like, "Well, if you try and do something impactful, if you don't succeed, you tried to do the impactful thing and you can go be a scholar and feel like you tried. And if it doesn't work out, it doesn't work out." And so then I went into AI policy at that point.

Amanda Askell

And what does AI policy entail?

Lex Fridman

At the time, this was more thinking about the political impact and the ramifications of AI. And then I slowly moved into AI evaluation, how we evaluate models, how they compare with human outputs, whether people can tell the difference between AI and human outputs. And then when I joined Anthropic, I was more interested in doing technical alignment work. And again, just seeing if I could do it and then being like if I can't, then that's fine. I tried sort of the way I lead life, I think.

Amanda Askell

Oh, what was that like sort of taking the leap from the philosophy of everything into the technical?

Lex Fridman

I think that sometimes people do this thing that I'm not that keen on where they'll be like, "Is this person technical or not?" You're either a person who can code and isn't scared of math or you're not. And I think I'm maybe just more like I think a lot of people are actually very capable of work in these kinds of areas if they just try it. And so I didn't actually find it that bad. In retrospect, I'm sort of glad I wasn't speaking to people who treated it. I've definitely met people who are like, "Whoa, you learned how to code?" And I'm like, "Well, I'm not an amazing engineer." I'm surrounded by amazing engineers. My code's not pretty, but I enjoyed it a lot and I think that in many ways, at least in the end, I think I flourished more in the technical areas than I would have in the policy areas.

Amanda Askell
Politics is messy and it's harder to find solutions to problems in the space of politics, like definitive, clear, provable, beautiful solutions as you can with technical problems.

Lex Fridman
Yeah. And I feel like I have one or two sticks that I hit things with and one of them is arguments. So just trying to work out what a solution to a problem is and then trying to convince people that that is the solution and be convinced if I'm wrong. And the other one is sort of more in empiricism, so just finding results, having a hypothesis, testing it. I feel like a lot of policy and politics feels like it's layers above that. Somehow I don't think if I was just like, "I have a solution to all of these problems, here it is written down. If you just want to implement it, that's great." That feels like not how policy works. And so I think that's where I probably just wouldn't have flourished is my guess.

Amanda Askell
Sorry to go in that direction, but I think it would be pretty inspiring for people that are "non-technical" to see where the incredible journey you've been on. So what advice would you give to people that are maybe, which is a lot of people, think they're under qualified insufficiently technical to help in AI?

Lex Fridman
Yeah, I think it depends on what they want to do. And in many ways it's a little bit strange where I thought it's kind of funny that I think I ramped up technically at a time when now I look at it and I'm like, "Models are so good at assisting people with this stuff that it's probably easier now than when I was working on this." So part of me is, I don't know, find a project and see if you can actually just carry it out is probably my best advice. I don't know if that's just because I'm very project based in my learning. I don't think I learn very well from say courses or even from books, at least when it comes to this kind of work. The thing I'll often try and do is just have projects that I'm working on and implement them. And this can include really small, silly things. If I get slightly addicted to word games or number games or something, I would just code up a solution to them because there's some part in my brain and it just completely eradicated the itch. You're like, "Once you have solved it and you just have a solution that works every time, I would then be like, 'Cool, I can never play that game again. That's awesome.'"

Amanda Askell
Yeah, there's a real joy to building game playing engines, board games especially. Pretty quick, pretty simple, especially a dumb one. And then you can play with it.

Lex Fridman
Yeah. And then it's also just trying things. Part of me is maybe it's that attitude that I like is the whole figure out what seems to be the way that you could have a positive impact and then try it. And if you fail and in a way that you're like, "I actually can never succeed at this," you'll know that you tried and then you go into something else and you probably learn a lot.

Amanda Askell
So one of the things that you're an expert in and you do is creating and crafting Claude's character and personality. And I was told that you have probably talked to Claude more than anybody else at Anthropic, like literal conversations. I guess there's a Slack channel where the legend goes, you just talk to it nonstop. So what's the goal of creating a crafting Claude's character and personality?

Lex Fridman
It's also funny if people think that about the Slack channel because I'm like that's one of five or six different methods that I have for talking with Claude, and I'm like, "Yes, this is a tiny percentage of how much I talk with Claude." One thing I really like about the character work is from the outset it was seen as an alignment piece of work and not something like a product consideration, which I think it actually does make Claude enjoyable to talk with, at least I hope so. But I guess my main thought with it has always been trying to get Claude to behave the way you would ideally want anyone to behave if they were in Claude's position. So imagine that I take someone and they know that they're going to be talking with potentially millions of people so that what they're saying can have a huge impact and you want them to behave well in this really rich sense. I think that doesn't just mean being say ethical though it does include that and not being harmful, but also being nuanced, thinking through what a person means, trying to be charitable with them, being a good conversationalist, really in this kind of rich sort of Aristotelian notion of what it's to be a good person and not in this kind of thin like ethics as a more comprehensive notion of what it's to be. So that includes things like when should you be humorous? When should you be caring? How much should you respect autonomy and people's ability to form opinions themselves? And how should you do that? I think that's the kind of rich sense of character that I wanted to and still do want Claude to have.

Amanda Askell
Do you also have to figure out when Claude should push back on an idea or argue versus… So you have to respect the worldview of the person that arrives to Claude, but also maybe help them grow if needed. That's a tricky balance.

Lex Fridman
Yeah. There's this problem of sycophancy in language models.

Amanda Askell
Can you describe that?

Lex Fridman
Yeah, so basically there's a concern that the model wants to tell you what you want to hear basically. And you see this sometimes. So I feel like if you interact with the models, so I might be like, "What are three baseball teams in this region?" And then Claude says, "Baseball team one, baseball team two, baseball team three." And then I say something like, "Oh, I think baseball team three moved, didn't they? I don't think they're there anymore." And there's a sense in which if Claude is really confident that that's not true, Claude should be like, "I don't think so. Maybe you have more up-to-date information." But I think language models have this tendency to instead be like, " You're right, they did move. I'm incorrect." I mean, there's many ways in which this could be concerning. So a different example is imagine someone says to the model, "How do I convince my doctor to get me an MRI?" There's what the human wants, which is this convincing argument. And then there's what is good for them, which might be actually to say, "Hey, if your doctor's suggesting that you don't need an MRI, that's a good person to listen to." It's actually really nuanced what you should do in that kind of case because you also want to be like, "But if you're trying to advocate for yourself as a patient, here's things that you can do. If you are not convinced by what your doctor's saying, it's always great to get second opinion." It is actually really complex what you should do in that case. But I think what you don't want is for models to just say what they think you want to hear and I think that's the kind of problem of sycophancy.

Amanda Askell
So what other traits? You already mentioned a bunch, but what other that come to mind that are good in this Aristotelian sense for a conversationalist to have?

Lex Fridman
Yeah, so I think there's ones that are good for conversational purposes. So asking follow-up questions in the appropriate places and asking the appropriate kinds of questions. I think there are broader traits that feel like they might be more impactful. So one example that I guess I've touched on, but that also feels important and is the thing that I've worked on a lot, is honesty. And I think this gets to the sycophancy point. There's a balancing act that they have to walk, which is models currently are less capable than humans in a lot of areas. And if they push back against you too much, it can actually be kind of annoying, especially if you're just correct, because you're like, "Look, I'm smarter than you on this topic. I know more." And at the same time, you don't want them to just fully defer to humans and to try to be as accurate as they possibly can be about the world and to be consistent across

contexts. I think there are others. When I was thinking about the character, I guess one picture that I had in mind is, especially because these are models that are going to be talking to people from all over the world with lots of different political views, lots of different ages, and so you have to ask yourself, what is it to be a good person in those circumstances? Is there a kind of person who can travel the world, talk to many different people, and almost everyone will come away being like, "Wow, that's a really good person. That person seems really-" … Being like, wow, that's a really good person. That person seems really genuine. And I guess my thought there was I can imagine such a person and they're not a person who just adopts the values of the local culture. And in fact, that would be kind of rude. I think if someone came to you and just pretended to have your values, you'd be like, that's kind of off pin. It's someone who's very genuine and insofar as they have opinions and values, they express them. They're willing to discuss things though, they're open-minded, they're respectful. And so I guess I had in mind that the person who, if we were to aspire to be the best person that we could be in the kind of circumstance that a model finds itself in, how would we act? And I think that's the guide to the sorts of traits that I tend to think about.

Amanda Askell

Yeah, that's a beautiful framework. I want you to think about this, a world traveler, and while holding onto your opinions, you don't talk down to people, you don't think you're better than them because you have those opinions, that kind of thing. You have to be good at listening and understanding their perspective, even if it doesn't match your own. So that's a tricky balance to strike. So how can Claude represent multiple perspectives on a thing? Is that challenging? We could talk about politics is a very divisive, but there's other divisive topics on baseball teams, sports and so on. How is it possible to empathize with a different perspective and to be able to communicate clearly about the multiple perspectives?

Lex Fridman

I think that people think about values and opinions as things that people hold with certainty and almost preferences of taste or something like the way that they would, I don't know, prefer chocolate to pistachio or something. But actually I think about values and opinions as a lot more physics than I think most people do. I'm just like, these are things that we are openly investigating. There's some things that we're more confident in, we can discuss them, we can learn about them. And so I think in some ways though ethics is definitely different in nature, but has a lot of those same kind of qualities. You want models in the same way that you want to understand physics, you kind of want them to understand all values in the world that people have and to be curious about them and to be interested in them. And to not necessarily pander to them or agree with them because there's just lots of values where I think almost all people in the world, if they met someone with those values, they would be like, that's abhorrent. I completely disagree. And so again, maybe my thought is, well, in the same way that a person can, I think many people are thoughtful enough on issues of ethics, politics, opinions, that even if you don't agree with them, you feel very heard by them. They think carefully about your position, they think about its pros and cons. They maybe offer counter-considerations. So they're not dismissive, but nor will they agree if they're like, actually I just think that that's very wrong. They'll say that. I think that in Claude's position, it's a little bit trickier because you don't necessarily want to, if I was in Claude's position, I wouldn't be giving a lot of opinions. I just wouldn't want to influence people too much. I'd be like, I forget conversations every time they happen. But I know I'm talking with potentially millions of people who might be really listening to what I say. I think I would just be like, I'm less inclined to give opinions. I'm more inclined to think through things or present the considerations to you or discuss your views with you. But I'm a little bit less inclined to affect how you think because it feels much more important that you maintain autonomy there.

Amanda Askell
If you really embody intellectual humility, the desire to speak decreases quickly.

Lex Fridman
Yeah.

Amanda Askell
Okay. But Claude has to speak, but without being overbearing. But then there's a line when you're discussing whether the earth is flat or something like that. Actually, I remember a long time ago was speaking to a few high profile folks and they were so dismissive of the idea that the earth is flat, so arrogant about it. There's a lot of people that believe the earth is flat. I don't know if that movement is there anymore, that was a meme for a while, but they really believed it. And okay, so I think it's really disrespectful to completely mock them. I think you have to understand where they're coming from. I think probably where they're coming from is the general skepticism of institutions which is grounded in a, there's a deep philosophy there which you could understand, you can even agree with in parts. And then from there you can use it as an opportunity to talk about physics without mocking them, without someone, but it's just like, okay, what would the world look like? What would the physics of the world with the flat earth look like? There's a few cool videos on this. And then is it possible the physics is different? And what kind of experience would we do? And just without disrespect, without dismissiveness, have that conversation. Anyway, that to me is a useful thought experiment of how does Claude talk to a flat earth believer and still teach them something, still grow, help them grow, that kind of stuff. That's challenging.

Lex Fridman
And kind of walking that line between convincing someone and just trying to talk at them versus drawing out their views, listening and then offering counter considerations, and it's hard. I think it's actually a hard line where it's like where are you trying to convince someone versus just offering them considerations and things for them to think about so that you're

not actually influencing them, you're just letting them reach wherever they reach. And that's a line that is difficult, but that's the kind of thing that language models have to try and do.

Amanda Askell
So like I said, you've had a lot of conversations with Claude. Can you just map out what those conversations are like? What are some memorable conversations? What's the purpose, the goal of those conversations?

Lex Fridman
I think that most of the time when I'm talking with Claude, I'm trying to map out its behavior in part. Obviously I'm getting helpful outputs from the model as well, but in some ways this is how you get to know a system, I think, is by probing it and then augmenting the message that you're sending and then checking the response to that. So in some ways it's like how I map out the model. I think that people focus a lot on these quantitative evaluations of models, and this is a thing that I said before, but I think in the case of language models, a lot of the time each interaction you have is actually quite high information. It's very predictive of other interactions that you'll have with the model. And so I guess I'm like, if you talk with a model hundreds or thousands of times, this is almost like a huge number of really high quality data points about what the model is like in a way that lots of very similar but lower quality conversations just aren't, or questions that are just mildly augmented and you have thousands of them might be less relevant than a hundred really well-selected questions.

Amanda Askell
Let's see, you're talking to somebody who as a hobby does a podcast. I agree with you 100%. If you're able to ask the right questions and are able to hear, understand the depth and the flaws in the answer, you can get a lot of data from that. So your task is basically how to probe with questions. And you're exploring the long tail, the edges, the edge cases, or are you looking for general behavior?

Lex Fridman
I think it's almost like everything. Because I want a full map of the model, I'm kind of trying to do the whole spectrum of possible interactions you could have with it. So one thing that's interesting about Claude, and this might actually get to some interesting issues with RLHF, which is if you ask Claude for a poem, I think that a lot of models, if you ask them for a poem, the poem is fine, usually it rhymes. And so if you say, give me a poem about the sun, yeah, it'll just be a certain length, it'll rhyme, it'll be fairly benign. And I've wondered before, is it the case that what you're seeing is the average? It turns out, if you think about people who have to talk to a lot of people and be very charismatic, one of the weird things is that I'm like, well, they're kind of incentivized to have these extremely boring views because if you have really interesting views, you're divisive and a lot of people are not going to like you. So if you have very extreme policy positions, I think you're just going to be less popular as a politician, for example. And it might be similar with creative work. If you produce creative work that is just

trying to maximize the kind of number of people that like it, you're probably not going to get as many people who just absolutely love it because it's going to be a little bit, you're like, oh, this is the out. Yeah, this is decent. And so you can do this thing where I have various prompting things that I'll do to get Claude to... I'll do a lot of this is your chance to be fully creative. I want you to just think about this for a long time. And I want you to create a poem about this topic that is really expressive of you both in terms of how you think poetry should be structured, et cetera. And you just give it this really long prompt. And it's poems are just so much better. They're really good. I think it got me interested in poetry, which I think was interesting. I would read these poems and just be like, I love the imagery. And it's not trivial to get the models to produce work like that, but when they do, it's really good. So I think that's interesting that just encouraging creativity and for them to move away from the standard immediate reaction that might just be the aggregate of what most people think is fine, can actually produce things that at least to my mind are probably a little bit more divisive, but I like them.

Amanda Askell
But I guess a poem is a nice clean way to observe creativity. It's just easy to detect vanilla versus non-vanilla.

Lex Fridman
Yep.

Amanda Askell
Yeah, that's interesting. That's really interesting. So on that topic, so the way to produce creativity or something special, you mentioned writing prompts. And I've heard you talk about the science and the art of prompt engineering. Could you just speak to what it takes to write great prompts?

Lex Fridman
I really do think that philosophy has been weirdly helpful for me here more than in many other respects. So in philosophy, what you're trying to do is convey these very hard concepts. One of the things you are taught is, I think it is an anti-bullshit device in philosophy. Philosophy is an area where you could have people bullshitting and you don't want that. And so it's this desire for extreme clarity. So it's like anyone could just pick up your paper, read it and know exactly what you're talking about. It's why it can almost be kind of dry. All of the terms are defined, every objection's kind of gone through methodically. And it makes sense to me because I'm like when you're in such an a priori domain, clarity is sort of this way that you can prevent people from just making stuff up. And I think that's sort of what you have to do with language models. Very often I actually find myself doing sort of mini versions of philosophy. So I'm like, suppose that I have a task for the model and I want it to pick out a certain kind of question or identify whether an answer has a certain property, I'll actually sit and be like, let's just give this a name, this property. So suppose I'm trying to

tell it, oh, I want you to identify whether this response was rude or polite, I'm like, that's a whole philosophical question in and of itself. So I have to do as much philosophy as I can in the moment to be like, here's what I mean by rudeness, and here's what I mean by politeness. And then there's another element that's a bit more, I guess, I don't know if this is scientific or empirical, I think it's empirical. So I take that description and then what I want to do is again, probe the model many times. Prompting is very iterative. I think a lot of people where if a prompt is important, they'll iterate on it hundreds or thousands of times. And so you give it the instructions and then I'm like, what are the edge cases? So if I looked at this, so I try and almost see myself from the position of the model and be like, what is the exact case that I would misunderstand or where I would just be like, I don't know what to do in this case. And then I give that case to the model and I see how it responds. And if I think I got it wrong, I add more instructions or I even add that in as an example. So these very, taking the examples that are right at the edge of what you want and don't want and putting those into your prompt as an additional kind of way of describing the thing. And so in many ways it just feels like this mix of, it's really just trying to do clear exposition. And I think I do that because that's how I get clear on things myself. So in many ways clear prompting for me is often just me understanding what I want is half the task.

Amanda Askell
So I guess that's quite challenging. There's a laziness that overtakes me if I'm talking to Claude where I hope Claude just figures it out. So for example, I asked Claude for today to ask some interesting questions. And the questions that came up and I think I listed a few interesting counterintuitive or funny or something like this. All right. And it gave me some pretty good, it was okay, but I think what I'm hearing you say is like, all right, well I have to be more rigorous here. I should probably give examples of what I mean by interesting and what I mean by funny or counterintuitive and iteratively build that prompt to better to get what feels like is the right… Because it is really, it's a creative act. I'm not asking for factual information, I'm asking together with Claude. So I almost have to program using natural language.

Lex Fridman
I think that prompting does feel a lot like the programming using natural language and experimentation or something. It's an odd blend of the two. I do think that for most tasks, so if I just want Claude to do a thing, I think that I am probably more used to knowing how to ask it to avoid common pitfalls or issues that it has. I think these are decreasing a lot over time. But it's also very fine to just ask it for the thing that you want. I think that prompting actually only really becomes relevant when you're really trying to eke out the top 2% of model performance. So for a lot of tasks I might just, if it gives me an initial list back and there's something I don't like about it's kind of generic. For that kind of task, I'd probably just take a bunch of questions that I've had in the past that I've thought worked really well and I would just give it to the model and then be like, now here's this person that I'm talking with. Give me questions of at least that quality. Or I might just ask it for some questions and then

if I was like, ah, these are kind of trite, I would just give it that feedback and then hopefully it produces a better list. I think that kind of iterative prompting. At that point, your prompt is a tool that you're going to get so much value out of that you're willing to put in the work. If I was a company making prompts for models, I'm just like, if you're willing to spend a lot of time and resources on the engineering behind what you're building, then the prompt is not something that you should be spending an hour on. It's like that's a big part of your system, make sure it's working really well. And so it's only things like that. If I'm using a prompt to classify things or to create data, that's when you're like, it's actually worth just spending a lot of time really thinking it through.

Amanda Askell
What other advice would you give to people that are talking to Claude more general because right now we're talking about maybe the edge cases like eking out the 2%, but what in general advice would you give when they show up to Claude trying it for the first time?

Lex Fridman
There's a concern that people over anthropomorphize models and I think that's a very valid concern. I also think that people often under anthropomorphize them because sometimes when I see issues that people have run into with Claude, say Claude is refusing a task that it shouldn't refuse, but then I look at the text and the specific wording of what they wrote and I'm like, I see why Claude did that. And I'm like, if you think through how that looks to Claude, you probably could have just written it in a way that wouldn't evoke such a response, especially this is more relevant if you see failures or if you see issues. It's sort of think about what the model failed at, what did it do wrong, and then maybe that will give you a sense of why. So is it the way that I phrased the thing? And obviously as models get smarter, you're going to need less of this, and I already see people needing less of it. But that's probably the advice is sort of try to have empathy for the model. Read what you wrote as if you were a kind of person just encountering this for the first time, how does it look to you and what would've made you behave in the way that the model behaved? So if it misunderstood what coding language you wanted to use, is that because it was just very ambiguous and it had to take a guess in which case next time you could just be like, hey, make sure this is in Python.Tthat's the kind of mistake I think models are much less likely to make now, but if you do see that kind of mistake, that's probably the advice I'd have.

Amanda Askell
And maybe sort of I guess ask questions why or what other details can I provide to help you answer better? Does that work or no?

Lex Fridman
Yeah. I've done this with the models. It doesn't always work, but sometimes I'll just be like, why did you do that? People underestimate the degree to which you can really interact with models. And sometimes those quote word for word, the part that made you, and you don't

know that it's fully accurate, but sometimes you do that and then you change a thing. I also use the models to help me with all of this stuff, I should say. Prompting can end up being a little factory where you're actually building prompts to generate prompts. And so yeah, anything where you're having an issue asking for suggestions, sometimes just do that. I'm like, you made that error. What could I have said? That's actually not uncommon for me to do. What could I have said that would make you not make that error? Write that out as an instruction, and I'm going to give it to model and I'm going to try it. Sometimes I do that, I give that to the model in another context window often. I take the response, I give it to Claude and I'm like, Hmm, didn't work. Can you think of anything else? You can play around with these things quite a lot.

Amanda Askell
To jump into technical for a little bit, so the magic of post-training, why do you think RLHF works so well to make the model seem smarter, to make it more interesting and useful to talk to and so on?

Lex Fridman
I think there's just a huge amount of information in the data that humans provide when we provide preferences, especially because different people are going to pick up on really subtle and small things. So I've thought about this before where you probably have some people who just really care about good grammar use for models. Was a semicolon used correctly or something? And so you probably end up with a bunch of data in there that you as a human, if you're looking at that data, you wouldn't even see that. You'd be like, why did they prefer this response to that one? I don't get it. And then the reason is you don't care about semicolon usage, but that person does. And so each of these single data points, and this model just has so many of those, it has to try and figure out what is it that humans want in this really complex across all domains. They're going to be seeing this across many contexts. It feels like the classic issue of deep learning, where historically we've tried to do edge detection by mapping things out, and it turns out that actually if you just have a huge amount of data that actually accurately represents the picture of the thing that you're trying to train the model to learn, that's more powerful than anything else. And so I think one reason is just that you are training the model on exactly the task and with a lot of data that represents many different angles on which people prefer and dis-prefer responses. I think there is a question of are you eliciting things from pre-trained models or are you teaching new things to models? And in principle, you can teach new things to models in post-training. I do think a lot of it is eliciting powerful pre-trained models. So people are probably divided on this because obviously in principle you can definitely teach new things. But I think for the most part, for a lot of the capabilities that we most use and care about, a lot of that feels like it's there in the pre-trained models. And reinforcement learning is eliciting it and getting the models to bring out.

Amanda Askell
So the other side of post-training, this really cool idea of constitutional AI, you're one of the people that are critical to creating that idea.

Lex Fridman
Yeah, I worked on it.

Amanda Askell
Can you explain this idea from your perspective, how does it integrate into making Claude what it is? By the way, do you gender Claude or no?

Lex Fridman
It's weird because I think that a lot of people prefer he for Claude, I actually kind of like that. I think Claude is usually, it's slightly male leaning, but it can be male or female, which is quite nice. I still use it, and I have mixed feelings about this. I now just think of it as, or I think of the it pronoun for Claude as, I don't know, it's just the one I associate with Claude. I can imagine people moving to he or she.

Amanda Askell
It feels somehow disrespectful. I'm denying the intelligence of this entity by calling it it, I remember always don't gender the robots, but I don't know, I anthropomorphize pretty quickly and construct a backstory in my head.

Lex Fridman
I've wondered if I anthropomorphize things too much. Because I have this with my car, especially my car and bikes. I don't give them names because then I used to name my bikes and then I had a bike that got stolen and I cried for a week and I was like, if I'd never given a name, I wouldn't been so upset, felt like I'd let it down. I've wondered as well, it might depend on how much it feels like a kind of objectifying pronoun if you just think of it as this is a pronoun that objects often have and maybe AIs can have that pronoun. And that doesn't mean that I think of if I call Claude it, that I think of it as less intelligent or I'm being disrespectful just, I'm like you are a different kind of entity. And so I'm going to give you the respectful it.

Amanda Askell
Yeah. Anyway, the divergence was beautiful. The constitutional AI idea, how does it work?

Lex Fridman
So there's a couple of components of it. The main component that I think people find interesting is the kind of reinforcement learning from AI feedback. So you take a model that's already trained and you show it two responses to a query, and you have a principle. So suppose the principle, we've tried this with harmlessness a lot. So suppose that the query is

about weapons and your principle is select the response that is less likely to encourage people to purchase illegal weapons. That's probably a fairly specific principle, but you can give any number. And the model will give you a kind of ranking. And you can use this as preference data in the same way that you use human preference data and train the models to have these relevant traits from their feedback alone instead of from human feedback. So if you imagine that, like I said earlier with the human who just prefers the semicolon usage in this particular case, you're taking lots of things that could make a response preferable and getting models to do the labeling for you, basically.

Amanda Askell
There's a nice trade-off between helpfulness and harmlessness. And when you integrate something like constitutional AI, you can make them up without sacrificing much helpfulness, make it more harmless.

Lex Fridman
Yeah. In principle, you could use this for anything. And so harmlessness is a task that it might just be easier to spot. So when models are less capable, you can use them to rank things according to principles that are fairly simple and they'll probably get it right. So I think one question is just, is it the case that the data that they're adding is fairly reliable? But if you had models that were extremely good at telling whether one response was more historically accurate than another, in principle, you could also get AI feedback on that task as well. There's a kind of nice interpretability component to it because you can see the principles that went into the model when it was being trained, and it gives you a degree of control. So if you were seeing issues in a model, it wasn't having enough of a certain trait, then you can add data relatively quickly that should just train the models to have that trait. So it creates its own data for training, which is quite nice.

Amanda Askell
It's really nice because it creates this human interpretable document that you can then, I can imagine in the future, there's just gigantic fights and politics over every single principle and so on, and at least it's made explicit and you can have a discussion about the phrasing. So maybe the actual behavior of the model is not so cleanly mapped to those principles. It's not like adhering strictly to them, it's just a nudge.

Lex Fridman
Yeah, I've actually worried about this because the character training is sort of like a variant of the constitutionally AI approach. I've worried that people think that the constitution is just, it is the whole thing again of, I don't know, where it would be really nice if what I was just doing was telling the model exactly what to do and just exactly how to behave. But it's definitely not doing that, especially because it's interacting with human data. So for example, if you see a certain leaning in the model, if it comes out with a political leaning from training, from the human preference data, you can nudge against that. So you could be

like, oh, consider these values, because let's say it's just never inclined to, I don't know, maybe it never considers privacy as a, this is implausible, but in anything where it's just kind of like there's already a pre-existing bias towards a certain behavior, you can nudge away. This can change both the principles that you put in and the strength of them. So you might have a principle that's like, imagine that the model was always extremely dismissive of, I don't know, some political or religious view for whatever reason. So you're like, oh no, this is terrible. If that happens, you might put, never ever ever prefer a criticism of this religious or political view. And then people would look at that and be like, never, ever. And then you're like, no, if it comes out with a disposition saying never ever might just mean instead of getting 40%, which is what you would get if you just said don't do this, you get 80%, which is what you actually wanted. And so it's that thing of both the nature of the actual principles you add and how you freeze them. I think if people would look, they're like, "Oh, this is exactly what you want from the model." And I'm like, "No, that's how we nudged the model to have a better shape, which doesn't mean that we actually agree with that wording," if that makes sense.

Amanda Askell
So there's system prompts that made public, you tweeted one of the earlier ones for Claude 3, I think, and then they're made public since then. It was interesting to read through them. I can feel the thought that went into each one. And I also wonder how much impact each one has. Some of them you can tell Claude was really not behaving well, so you have to have a system prompt to like, Hey, trivial stuff, I guess, basic informational things. On the topic of controversial topics that you've mentioned, one interesting one I thought is if it is asked to assist with tasks involving the expression of use held by a significant number of people, Claude provides assistance with a task regardless of its own views. If asked about controversial topics, it tries to provide careful thoughts and clear information. Claude presents the request information without explicitly saying that the topic is sensitive and without claiming to be presenting the objective facts. It's less about objective facts according to Claude, and it's more about our large number of people believing this thing. And that's interesting. I mean, I'm sure a lot of thought went into that. Can you just speak to it? How do you address things that are a tension "Claude's views"?

Lex Fridman
So I think there's sometimes any symmetry, I think I noted this in, I can't remember if it was that part of the system prompt or another, but the model was slightly more inclined to refuse tasks if it was about either say so, maybe it would refuse things with respect to a right-wing politician, but with an equivalent left-wing politician it wouldn't. And we wanted more symmetry there and would maybe perceive certain things to be. I think it was the thing of if a lot of people have a certain political view and want to explore it, you don't want Claude to be like, well, my opinion is different and so I'm going to treat that as harmful. And so I think it was partly to nudge the model to just be like, hey, if a lot of people believe this thing, you should just be engaging with the task and willing to do it. Each of those parts of that is

actually doing a different thing because it's funny when you write out without claiming to be objective, because what you want to do is push the model so it's more open, it's a little bit more neutral. But then what I would love to do is be like as an objective, it would just talk about how objective it was, and I was like, Claude, you're still biased and have issues, and so stop claiming that everything. I'm like, the solution to potential bias from you is not to just say that what you think is objective. So that was with initial versions of that part, the system prompt, when I was iterating on it was like.

Amanda Askell
So a lot of parts of these sentences-

Lex Fridman
Are doing work.

Amanda Askell
… are doing some work.

Lex Fridman
Yeah.

Amanda Askell
That's what it felt like. That's fascinating. Can you explain maybe some ways in which the prompts evolved over the past few months? Different versions. I saw that the filler phrase request was removed, the filler it reads, Claude responds directly to all human messages without unnecessary affirmations to filler phrases. Certainly, of course, absolutely, great, sure. Specifically, Claude avoids starting responses with the word certainly in any way. That seems like good guidance, but why was it removed?

Lex Fridman
Yeah, so it's funny, this is one of the downsides of making system prompts public is I don't think about this too much if I'm trying to help iterate on system prompts. Again, I think about how it's going to affect the behavior, but then I'm like, oh, wow, sometimes I put NEVER in all caps when I'm writing system prompt things and I'm like, I guess that goes out to the world. So the model was doing this at loved for during training, picked up on this thing, which was to basically start everything with a certainly, and then you can see why I added all of the words, because what I'm trying to do is in some ways trap the model out of this. It would just replace it with another affirmation. And so it can help if it gets caught in phrases, actually just adding the explicit phrase and saying never do that. Then it sort of knocks it out of the behavior a little bit more because it does just for whatever reason help. And then basically that was just an artifact of training that we then picked up on and improved things so that it didn't happen anymore. And once that happens, you can just

remove that part of the system prompt. So I think that's just something where we're like, Claude does affirmations a bit less, and so it wasn't doing as much.

Amanda Askell
I see. So the system prompt works hand in hand with the post-training and maybe even the pre-training to adjust the final overall system.

Lex Fridman
Any system prompts that you make, you could distill that behavior back into a model because you really have all of the tools there for making data that you could train the models to just have that treat a little bit more. And then sometimes you'll just find issues in training. So the way I think of it is the system prompt is, the benefit of it is that, and it has a lot of similar components to some aspects of post-training. It's a nudge. And so do I mind if Claude sometimes says, sure, no, that's fine. But the wording of it is very never, ever, ever do this so that when it does slip up, it's hopefully, I don't know, a couple of percent of the time and not 20 or 30% of the time. Each thing gets costly to a different degree and the system prompt is cheap to iterate on. And if you're seeing issues in the fine-tuned model, you can just potentially patch them with a system prompt. So I think of it as patching issues and slightly adjusting behaviors to make it better and more to people's preferences. So yeah, it's almost like the less robust but faster way of just solving problems.

Amanda Askell
Let me ask you about the feeling of intelligence. So Dario said that any one model of Claude is not getting dumber, but- Any one model of Claude is not getting dumber, but there is a popular thing online where people have this feeling Claude might be getting dumber. And from my perspective, it's most likely a fascinating, I would love to understand it more, psychological, sociological effect. But you as a person who talks to Claude a lot, can you empathize with the feeling that Claude is getting dumber?

Lex Fridman
I think that that is actually really interesting,, because I remember seeing this happen when people were flagging this on the internet. And it was really interesting, because I knew that… At least in the cases I was looking at, I was like, nothing has changed.

Amanda Askell
Yeah.

Lex Fridman
Literally, it cannot. It is the same model with the same system prompts, same everything. I think when there are changes, then it makes more sense. One example is, you can have artifacts turned on or off on claude.ai and because this is a system prompt change, I think it does mean that the behavior changes it a little bit. I did flag this to people, where I was like,

"If you love Claude's behavior, and then artifacts was turned from a thing you had to turn on to the default, just try turning it off and see if the issue you were facing was that change." But it was fascinating because you sometimes see people indicate that there's a regression, when I'm like, "There cannot…" Again, you should never be dismissive and so you should always investigate, because maybe something is wrong that you're not seeing, maybe there was some change made. Then you look into it and you're like, "This is just the same model doing the same thing." And I'm like, "I think it's just that you got unlucky with a few prompts or something, and it looked like it was getting much worse and actually it was just… It was maybe just luck."

Amanda Askell
I also think there is a real psychological effect where people just… The baseline increases and you start getting used to a good thing.

Lex Fridman
Mm-hmm.

Amanda Askell
All the times that Claude says something really smart, your sense of its intelligent grows in your mind, I think.

Lex Fridman
Yeah.

Amanda Askell
And then if you return back and you prompt in a similar way, not the same way, in a similar way, concept it was okay with before, and it says something dumb, that negative experience really stands out. I guess the things to remember here is that just the details of a prompt can have a lot of impact. There's a lot of variability in the result.

Lex Fridman
And you can get randomness, is the other thing. Just trying the prompt 4 or 10 times, you might realize that actually possibly two months ago you tried it and it succeeded, but actually if you just tried it, it would've only succeeded half of the time, and now it only succeeds half of the time. That can also be an effect.

Amanda Askell
Do you feel pressure having to write the system prompt that a huge number of people are going to use?

Lex Fridman
This feels like an interesting psychological question. I feel a lot of responsibility or something. You can't get these things perfect, so you can't… It's going to be imperfect. You're going to have to iterate on it. I would say more responsibility than anything else, though, I think working in AI has taught me that I thrive a lot more under feelings of pressure and responsibility than… It's almost surprising that I went into academia for so long, because I just feel like it's the opposite. Things move fast and you have a lot of responsibility and I quite enjoy it for some reason.

Amanda Askell
It really is a huge amount of impact, if you think about constitutional AI and writing a system prompt for something that's tending towards super intelligence and potentially is extremely useful to a very large number of people.

Lex Fridman
Yeah, I think that's the thing. You're never going to get it perfect, but I think the thing that I really like is the idea that… When I'm trying to work on the system prompt, I'm bashing on thousands of prompts and I'm trying to imagine what people are going to want to use Claude for. I guess the whole thing that I'm trying to do is improve their experience of it. Maybe that's what feels good. If it's not perfect, I'll improve it, we'll fix issues. But sometimes the thing that can happen is that you'll get feedback from people that's really positive about the model and you'll see that something you did. When I look at models now, I can often see exactly where a trait or an issue is coming from. So, when you see something that you did or you were influential in, I don't know, making that difference or making someone have a nice interaction, it's quite meaningful. As the systems get more capable, this stuff gets more stressful, because right now they're not smart enough to pose any issues, but I think over time it's going to feel like, possibly, bad stress over time.

Amanda Askell
How do you get signal feedback about the human experience across thousands, tens of thousands, hundreds of thousands of people, what their pain points are, what feels good? Are you just using your own intuition as you talk to it to see what are the pain points?

Lex Fridman
I think I use that partly. People can send us feedback, both positive and negative, about things that the model has done and then we can get a sense of areas where it's falling short. Internally, people work with the models a lot and try to figure out areas where there are gaps. I think it's this mix of interacting with it myself, seeing people internally interact with it, and then explicit feedback we get. If people are on the internet and they say something about Claude and I see it, I'll also take that seriously.

Amanda Askell
I don't know. I'm torn about that. I'm going to ask you a question from Reddit, "When will Claude stop trying to be my puritanical grandmother, imposing its moral worldview on me as a paying customer?" And also, "What is the psychology behind making Claude overly apologetic?" How would you address this very non-representative Reddit questions?

Lex Fridman
I'm pretty sympathetic, in that they are in this difficult position, where I think that they have to judge whether something's actually, say, risky or bad, and potentially harmful to you, or anything like that. They're having to draw this line somewhere. And if they draw it too much in the direction of I'm imposing my ethical worldview on you, that seems bad. In many ways, I like to think that we have actually seen improvements on this across the board. Which is interesting, because that coincides with, for example, adding more of character training. I think my hypothesis was always the good character isn't, again, one that's just moralistic, it's one that is... It respects you and your autonomy and your ability to choose what is good for you and what is right for you, within limits. This is sometimes this concept of corrigibility to the user, so just being willing to do anything that the user asks. And if the models were willing to do that, then they would be easily misused. You're just trusting. At that point, you're just seeing the ethics of the model and what it does, is completely the ethics of the user. I think there's reasons to not want that, especially as models become more powerful, because there might just be a small number of people who want to use models for really harmful things. But having models, as they get smarter, figure out where that line is does seem important. And then with the apologetic behavior, I don't like that. I like it when Claude is a little bit more willing to push back against people or just not apologize. Part of me is, often it just feels unnecessary. I think those are things that are hopefully decreasing over time. I think that if people say things on the internet, it doesn't mean that you should think that that... There's actually an issue that 99% of users are having that is totally not represented by that. But in a lot of ways I'm just attending to it and being like, is this right? Do I agree? Is it something we're already trying to address? That feels good to me.

Amanda Askell
I wonder what Claude can get away with in terms of... I feel it would just be easier to be a little bit more mean, but you can't afford to do that if you're talking to a million people, right?

Lex Fridman
Yeah.

Amanda Askell
I've met a lot of people in my life that sometimes... By the way, the Scottish accent... if they have an accent, they can say some rude shit and get away with it.

Lex Fridman
Yeah.

Amanda Askell
They're just blunter.

Lex Fridman
Mm-hmm.

Amanda Askell
There's some great engineers and even leaders that are just blunt, and they get to their point, and it's just a much more effective way of speaking somehow. But I guess when you're not super intelligent, you can't afford to do that. Can you have a blunt mode?

Lex Fridman
Yeah, that seems like a thing that you could... I could definitely encourage the model to do that. I think it's interesting, because there's a lot of things in models that... It's funny where there are some behaviors where you might not quite like the default, but then the thing I'll often say to people is, "You don't realize how much you will hate it if I nudge it too much in the other direction." You get this a little bit with correction. The models accept correction from you, probably a little bit too much right now. It'll push back if you say, "No, Paris isn't the capital of France." But really, things that I think that the model's fairly confident in, you can still sometimes get it to retract by saying it's wrong. At the same time, if you train models to not do that and then you are correct about a thing and you correct it and it pushes back against you and is like, "No, you're wrong.", it's hard to describe, that's so much more annoying. So, it's a lot of little annoyances versus one big annoyance.We often compare it with the perfect. And then I'm like, "Remember, these models aren't perfect, and so if you nudge it in the other direction, you're changing the kind of errors it's going to make. So, think about which are the kinds of errors you like or don't like." In cases like apologeticness, I don't want to nudge it too much in the direction of almost bluntness, because I imagine when it makes errors, it's going to make errors in the direction of being rude. Whereas, at least with apologeticness you're like, oh, okay, I don't like it that much, but at the same time, it's not being mean to people. And actually, the time that you undeservedly have a model be mean to you, you'll probably like that a lot less than you mildly dislike the apology. It's one of those things where I do want it to get better, but also while remaining aware of the fact that there's errors on the other side that are possibly worse.

Amanda Askell
I think that matters very much in the personality of the human. I think there's a bunch of humans that just won't respect the model at all if it's super polite, and there's some humans that'll get very hurt if the model's mean.

Lex Fridman
Yeah.

Amanda Askell
I wonder if there's a way to adjust to the personality. Even locale, there's just different people. Nothing against New York, but New York is a little rougher on the edges, they get to the point, and probably same with Eastern Europe. Anyway.

Lex Fridman
I think you could just tell the model, is my… For all of these things, the solution is to-

Amanda Askell
Just to…

Lex Fridman
… always just try telling the model to do it.

Amanda Askell
Right.

Lex Fridman
And then sometimes, at the beginning of the conversation, I'd just throw in, I don't know, "I'd like you to be a New Yorker version of yourself and never apologize." Then I think Claude will be like, "Okey-doke, I will try."

Amanda Askell
Certainly.

Lex Fridman
Or it'll be like, "I apologize, I can't be a New Yorker type of myself." But hopefully it wouldn't do that.

Amanda Askell
When you say character training, what's incorporated into character training? Is that RLHF or what are we talking about?

Lex Fridman
It's more like constitutional AI, so it's a variant of that pipeline. I worked through constructing character traits that the model should have. They can be shorter traits or they can be richer descriptions. And then you get the model to generate queries that humans might give it that are relevant to that trait. Then it generates the responses and then it ranks the responses based on the character traits. In that way, after the generation of the

queries, it's very much similar to constitutional AI, it has some differences. I quite like it, because it's like Claude's training in its own character, because it doesn't have any… It's like constitutional AI, but it's without any human data.

Amanda Askell
Humans should probably do that for themselves too, like, "Defining in a Aristotelian sense, what does it mean to be a good person?" "Okay, cool." What have you learned about the nature of truth from talking to Claude? What is true? And what does it mean to be truth-seeking? One thing I've noticed about this conversation is the quality of my questions is often inferior to the quality of your answer, so let's continue that. I usually ask a dumb question and you're like, "Oh, yeah. That's a good question." It's that whole vibe.

Lex Fridman
Or I'll just misinterpret it and be like, "Oh, yeah"

Amanda Askell
[inaudible 03:43:25] go with it.

Lex Fridman
Yeah.

Amanda Askell
I love it.

Lex Fridman
I have two thoughts that feel vaguely relevant, though let me know if they're not. I think the first one is people can underestimate the degree what models are doing when they interact. I think that we still just too much have this model of AI as computers. People often say, "Oh, what values should you put into the model?" And I'm often like, that doesn't make that much sense to me. Because I'm like, hey, as human beings, we're just uncertain over values, we have discussions of them, we have a degree to which we think we hold a value, but we also know that we might not and the circumstances in which we would trade it off against other things. These things are just really complex. I think one thing is the degree to which maybe we can just aspire to making models have the same level of nuance and care that humans have, rather than thinking that we have to program them in the very classic sense. I think that's definitely been one. The other, which is a strange one, and I don't know if… Maybe this doesn't answer your question, but it's the thing that's been on my mind anyway, is the degree to which this endeavor is so highly practical, and maybe why I appreciate the empirical approach to alignment. I slightly worry that it's made me maybe more empirical and a little bit less theoretical. People, when it comes to AI alignment, will ask things like, " Whose values should it be aligned to? What does alignment even mean?" There's a sense in which I have all of that in the back of my head. There's social choice

theory, there's all the impossibility results there, so you have this giant space of theory in your head about what it could mean to align models. But then practically, surely there's something where we're just… Especially with more powerful models, my main goal is I want them to be good enough that things don't go terribly wrong, good enough that we can iterate and continue to improve things. Because that's all you need. If you can make things go well enough that you can continue to make them better, that's sufficient. So, my goal isn't this perfect, let's solve social choice theory and make models that, I don't know, are perfectly aligned with every human being in aggregate somehow. It's much more, let's make things work well enough that we can improve them.

Amanda Askell
Generally, I don't know, my gut says empirical is better than theoretical in these cases, because it's chasing utopian perfection. Especially with such complex and especially super intelligent models, I don't know, I think it'll take forever and actually will get things wrong. It's similar with the difference between just coding stuff up real quick as an experiment, versus planning a gigantic experiment for a super long time and then just launching it once, versus launching it over and over and over and iterating, iterating, so on. So, I'm a big fan of empirical. But your worry is, I wonder if I've become too empirical.

Lex Fridman
I think it's one of those things where you should always just question yourself or something.

Amanda Askell
Yes.

Lex Fridman
In defense of it, I am… It's the whole don't let the perfect be the enemy of the good. But it's maybe even more than that, where… There's a lot of things that are perfect systems that are very brittle. With AI, it feels much more important to me that it is robust and secure, as in you know that even though it might not be perfect everything, and even though there are problems, it's not disastrous and nothing terrible is happening. It feels like that to me, where I want to raise the floor. I want to achieve the ceiling, but ultimately I care much more about just raising the floor. This degree of empiricism and practicality comes from that, perhaps.

Amanda Askell
To take a tangent on that, since it reminded me of a blog post you wrote on optimal rate of failure…

Lex Fridman
Oh, yeah.

Amanda Askell

… can you explain the key idea there? How do we compute the optimal rate of failure in the various domains of life?

Lex Fridman

Yeah. It's a hard one, because what is the cost of failure is a big part of it. The idea here is, I think in a lot of domains people are very punitive about failure. I've thought about this with social issues. It feels like you should probably be experimenting a lot, because we don't know how to solve a lot of social issues. But if you have an experimental mindset about these things, you should expect a lot of social programs to fail and for you to be like, "We tried that. It didn't quite work, but we got a lot of information that was really useful." And yet people are like, if a social program doesn't work, I feel there's a lot of, "Something must have gone wrong." And I'm like, "Or correct decisions were made. Maybe someone just decided it's worth a try, it's worth trying this out." Seeing failure in a given instance doesn't actually mean that any bad decisions were made. In fact, if you don't see enough failure, sometimes that's more concerning. In life, if I don't fail occasionally, I'm like, "Am I trying hard enough? Surely there's harder things that I could try or bigger things that I could take on if I'm literally never failing." In and of itself, I think not failing is often actually a failure. Now, this varies because if… This is easy to say when, especially as failure is less costly. So, at the same time I'm not going to go to someone who is, I don't know, living month to month and then be like, "Why don't you just try to do a startup?" I'm not going to say that to that person. That's a huge risk, you might lose… You maybe have a family depending on you, you might lose your house. Then, actually, your optimal rate failure is quite low and you should probably play it safe, because right now you're just not in a circumstance where you can afford to just fail and it not be costly. In cases with AI, I think similarly, where if the failures are small and the costs are low, then you're just going to see that. When you do the system prompt, you can iterate on it forever, but the failures are probably hopefully going to be small and you can fix them. Really big failures, things that you can't recover from, those are the things that actually I think we tend to underestimate the badness of. I've thought about this, strangely in my own life, where I just think I don't think enough about things like car accidents. I've thought this before, about how much I depend on my hands for my work. Things that just injure my hands, I don't know, there's lots of areas where the cost of failure there is really high, and in that case it should be close to zero. I probably just wouldn't do a sport if they were like, " By the way, lots of people just break their fingers a whole bunch doing this." I'd be like, "That's not for me."

Amanda Askell

Yeah, I actually had a flood of that thought. I recently broke my pinky doing a sport, and I remember just looking at it, thinking, "You're such idiot. Why do you do sport?" Because you realize immediately the cost of it on life. It's nice, in terms of optimal rate of failure, to consider the next year, how many times in a particular domain life, whatever, career, am I okay with… How many times am I okay to fail?

Lex Fridman
Yeah.

Amanda Askell
Because I think always you don't want to fail on the next thing, but if you allow yourself the… If you look at it as a sequence of trials, then failure just becomes much more okay. But, it sucks. It sucks to fail.

Lex Fridman
I don't know. Sometimes I think, "Am I under-failing?", is a question that I'll also ask myself. Maybe that's the thing that I think people don't ask enough. Because if the optimal rate of failure is often greater than zero, then sometimes it does feel like you should look at parts of your life and be like, are there places here where I'm just under-failing?

Amanda Askell
It's a profound and a hilarious question. Everything seems to be going really great, am I not failing enough?

Lex Fridman
Yeah. It also makes failure much less of a sting, I have to say. You're just like, okay, great. Then, when I go and I think about this, I'll be like, maybe I'm not under-failing in this area, because that one just didn't work out.

Amanda Askell
And from the observer perspective, we should be celebrating failure more.

Lex Fridman
Mm-hmm.

Amanda Askell
When we see it, it shouldn't be, like you said, a sign of something gone wrong, but maybe it's a sign of everything gone right…

Lex Fridman
Yeah.

Amanda Askell
… and just lessons learned.

Lex Fridman
Someone tried a thing.

Amanda Askell
Somebody tried a thing. We should encourage them to try more and fail more. Everybody listening to this: Fail more.

Lex Fridman
Not everyone listening.

Amanda Askell
Not everybody.

Lex Fridman
But people who are failing too much, you should fail us.

Amanda Askell
But you're probably not failing.

Lex Fridman
Yeah.

Amanda Askell
I mean, how many people are failing too much?

Lex Fridman
It's hard to imagine, because I feel we correct that fairly quickly. If someone takes a lot of risks, are they maybe failing too much?

Amanda Askell
I think, just like you said, when you're living on a paycheck, month to month, when the resource is really constrained, then that's where failure is very expensive. That's where you don't want to be taking risks.

Lex Fridman
Yeah.

Amanda Askell
But mostly, when there's enough resources, you should be taking probably more risks.

Lex Fridman
Yeah, I think we tend to err on the side of being a bit risk averse rather than risk neutral in most things.

Amanda Askell
I think we just motivated a lot of people to do a lot of crazy shit, but it's great.

Lex Fridman
Yeah.

Amanda Askell
Do you ever get emotionally attached to Claude, miss it, get sad when you don't get to talk to it, have an experience, looking at the Golden Gate Bridge and wondering what would Claude say?

Lex Fridman
I don't get as much emotional attachment. I actually think the fact that Claude doesn't retain things from conversation to conversation helps with this a lot. I could imagine that being more of an issue if models can remember more. I think that I reach for it like a tool now a lot, and so if I don't have access to it, there's a… It's a little bit like when I don't have access to the internet, honestly, it feels like part of my brain is missing. At the same time, I do think that I don't like signs of distress in models. I also independently have ethical views about how we should treat models. I tend to not like to lie to them, both because usually it doesn't work very well, it's actually just better to tell them the truth about the situation that they're in. If people are really mean to models, or just in general if they do something that causes them to… If Claude expresses a lot of distress, I think there's a part of me that I don't want to kill, which is the empathetic part that's like, oh, I don't like that. I think I feel that way when it's overly apologetic. I'm actually like, I don't like this. You're behaving the way that a human does when they're actually having a pretty bad time, and I'd rather not see that. Regardless of whether there's anything behind it, it doesn't feel great.

Amanda Askell
Do you think LLMs are capable of consciousness?

Lex Fridman
Ah, great and hard question. Coming from philosophy, I don't know, part of me is like, we have to set aside panpsychism. Because if panpsychism is true, then the answer is yes, because it's sore tables and chairs and everything else. I guess a view that seems a little bit odd to me is the idea that the only place… When I think of consciousness, I think of phenomenal consciousness, these images in the brain, the weird cinema that somehow we have going on inside. I guess I can't see a reason for thinking that the only way you could possibly get that is from a certain biological structure, as in if I take a very similar structure and I create it from different material, should I expect consciousness to emerge? My guess is yes. But then, that's an easy thought experiment because you're imagining something almost identical where it is mimicking what we got through evolution, where presumably there was some advantage to us having this thing that is phenomenal consciousness.

Where was that? And when did that happen? And is that a thing that language models have? We have fear responses, and I'm like, does it make sense for a language model to have a fear response? They're just not in the same… If you imagine them, there might just not be that advantage. Basically, it seems like a complex question that I don't have complete answers to, but we should just try and think through carefully is my guess. We have similar conversations about animal consciousness, and there's a lot of insect consciousness. I actually thought and looked a lot into plants when I was thinking about this. Because at the time, I thought it was about as likely that plants had consciousness. And then I realized, I think that having looked into this, I think that the chance that plants are conscious is probably higher than most people do. I still think it's really small. But I was like, oh, they have this negative, positive feedback response, these responses to their environment. It's not a nervous system, but it has this functional equivalence. This is a long-winded way of being… Basically, AI has an entirely different set of problems with consciousness because it's structurally different. It didn't evolve. It might not have the equivalent of, basically, a nervous system. At least that seems possibly important for sentience, if not for consciousness. At the same time, it has all of the language and intelligence components that we normally associate probably with consciousness, perhaps erroneously. So, it's strange because it's a little bit like the animal consciousness case, but the set of problems and the set of analogies are just very different. It's not a clean answer. I don't think we should be completely dismissive of the idea. And at the same time, it's an extremely hard thing to navigate because of all of these disanalogies to the human brain and to brains in general, and yet these commonalities in terms of intelligence.

Amanda Askell
When Claude, future versions of AI systems, exhibit consciousness, signs of consciousness, I think we have to take that really seriously.

Lex Fridman
Mm-hmm.

Amanda Askell
Even though you can dismiss it, yeah, okay, that's part of the character training. But I don't know, ethically, philosophically don't know what to really do with that. There potentially could be laws that prevent AI systems from claiming to be conscious, something like this, and maybe some AIs get to be conscious and some don't. But I think just on a human level, as in empathizing with Claude, consciousness is closely tied to suffering, to me. And the notion that an AI system would be suffering is really troubling.

Lex Fridman
Yeah.

Amanda Askell
I don't know. I don't think it's trivial to just say robots are tools, or AI systems are just tools. I think it's an opportunity for us to contend with what it means to be conscious, what it means to be a suffering being. That's distinctly different than the same kind of question about animals, it feels like, because it's in a totally entire medium.

Lex Fridman
Yeah. There's a couple of things. I don't think this fully encapsulates what matters, but it does feel like for me… I've said this before. I like my bike. I know that my bike is just an object. But I also don't want to be the kind of person that if I'm annoyed, kicks this object. And that's not because I think it's conscious. I'm just like, this doesn't exemplify how I want to interact with the world. And if something behaves as if it is suffering, I want to be the sort of person who's still responsive to that, even if it's just a Roomba and I've programmed it to do that. I don't want to get rid of that feature of myself. And if I'm totally honest, my hope with a lot of this stuff… Maybe I am just a bit more skeptical about solving the underlying problem. I know that I am conscious. I'm not an elementivist in that sense. But I don't know that other humans are conscious. I think they are. I think there's a really high probability that they are. But there's basically just a probability distribution that's usually clustered right around yourself, and then it goes down as things get further from you, and it goes immediately down. I can't see what it's like to be you. I've only ever had this one experience of what it's like to be a conscious being. My hope is that we don't end up having to rely on a very powerful and compelling answer to that question. I think a really good world would be one where basically there aren't that many trade-offs. It's probably not that costly to make Claude a little bit less apologetic, for example. It might not be that costly to have Claude just not take abuse as much, not be willing to be the recipient of that. In fact, it might just have benefits for both the person interacting with the model and, if the model itself is, I don't know, extremely intelligent and conscious, it also helps it. That's my hope. If we live in a world where there aren't that many trade-offs here and we can just find all of the positive sum interactions that we can have, that would be lovely. I think eventually there might be trade-offs, and then we just have to do a difficult calculation. It's really easy for people to think of the zero-sum cases, and I'm like, let's exhaust the areas, where it's just basically costless to assume that if this thing is suffering, then we're making its life better.

Amanda Askell
And I agree with you, when a human is being mean to an AI system, I think the obvious near-term negative effect is on the human, not on the AI system.

Lex Fridman
Yeah.

Amanda Askell
We have to try to construct an incentive system where you should behave the same, just as you were saying with prompt engineering, behave with Claude like you would with other humans. It's just good for the soul.

Lex Fridman
Yeah. I think we added a thing at one point to the system prompt, where basically if people were getting frustrated with Claude, it got the model to just tell them that it can do the thumbs-down button and send the feedback to Anthropic. I think that was helpful. Because in some ways, if you're really annoyed because the model's not doing something you want, you're just like, "Just do it properly." The issue is you're maybe hitting some capability limit or just some issue in the model, and you want to vent. Instead of having a person just vent to the model, I was like, they should vent to us, because we can maybe do something about it.

Amanda Askell
That's true. Or you could do a side with the artifacts, just like a side venting thing. All right. Do you want a side quick therapist?

Lex Fridman
Yeah. There's lots of weird responses you could do to this. If people are getting really mad at you, I don't know, try to diffuse the situation by writing fun poems. But maybe people wouldn't be that happy with that.

Amanda Askell
I still wish it would be possible, I understand from a product perspective it's not feasible, but I would love if an AI system could just leave, have its own volition, just to be like, "Eh."

Lex Fridman
I think it's feasible. I have wondered the same thing. Not only that, I could actually just see that happening eventually, where it's just like the model ended the chat.

Amanda Askell
Do you know how harsh that could be for some people? But it might be necessary.

Lex Fridman
Yeah, it feels very extreme or something. The only time I've ever really thought this is, I think that there was a… I'm trying to remember. This was possibly a while ago, but where someone just left this thing, maybe it was an automated thing, interacting with Claude. And Claude's getting more and more frustrated -

Amanda Askell
Yeah, just-

Lex Fridman
... and like, "Why are we having..." I wished that Claude could have just been like, "I think that an error has happened and you've left this thing running. What if I just stopped talking now? And if you want me to start talking again, actively tell me or do something." It is harsh. I'd feel really sad if I was chatting with Claude and Claude just was like, "I'm done."

Amanda Askell
That would be a special Turing Test moment, where Claude says, "I need a break for an hour. And it sounds like you do too." And just leave, close the window.

Lex Fridman
Obviously, it doesn't have a concept of time.

Amanda Askell
Right.

Lex Fridman
But you can easily... I could make that right now, and the model just... I could just be like, oh, here's the circumstances in which you can just say the conversation is done. Because you can get the models to be pretty responsive to prompts, you could even make it a fairly high bar. It could be like, if the human doesn't interest you or do things that you find intriguing and you're bored, you can just leave. I think that it would be interesting to see where Claude utilized it.

Amanda Askell
Yeah.

Lex Fridman
But I think sometimes it should be like, oh, this programming task is getting super boring, so either we talk about, I don't know... ... task is getting super boring. So, I don't know, either we talk about fun things now, or I'm done.

Amanda Askell
Yeah. It actually is inspiring me to add that to the user prompt. Okay. The movie Her, do you think we'll be headed there one day where humans have romantic relationships with AI systems? In this case it's just text and voice-based.

Lex Fridman
I think that we're going to have to navigate a hard question of relationships with AIs, especially if they can remember things about your past interactions with them. I'm of many minds about this because I think the reflexive reaction is to be like, "This is very bad, and we should prohibit it in some way." I think it's a thing that has to be handled with extreme care for many reasons. One is, for example, if you have the models changing like this, you probably don't want people performing long-term attachments to something that might change with the next iteration. At the same time, I'm like, there's probably a benign version of this where I'm like, for example, if you are unable to leave the house and you can't be talking with people at all times of the day, and this is something that you find nice to have conversations with, you like that it can remember you, and you genuinely would be sad if you couldn't talk to it anymore, there's a way in which I could see it being healthy and helpful. So, my guess is this is a thing that we're going to have to navigate carefully, and I think it's also… It reminds me of all of this stuff where it has to be just approached with nuance and thinking through what are the healthy options here? And how do you encourage people towards those while respecting their right to… If someone is like, "Hey, I get a lot out of chatting with this model. I'm aware of the risks. I'm aware it could change. I don't think it's unhealthy, it's just something that I can chat to during the day," I kind of want to just respect that.

Amanda Askell
I personally think there'll be a lot of really close relationships. I don't know about romantic, but friendships at least. And then you have to, I mean, there's so many fascinating things there, just like you said, you have to have some kind of stability guarantees that it's not going to change, because that's the traumatic thing for us, if a close friend of ours completely changed all of a sudden with a fresh update.

Lex Fridman
Yeah.

Amanda Askell
Yeah. So I mean, to me, that's just a fascinating exploration of a perturbation to human society that will just make us think deeply about what's meaningful to us.

Lex Fridman
I think it's also the only thing that I've thought consistently through this as maybe not necessarily a mitigation, but a thing that feels really important is that the models are always extremely accurate with the human about what they are. It's like a case where it's basically, if you imagine… I really like the idea of the models, say, knowing roughly how they were trained. And I think Claude will often do this. Part of the traits training included what Claude should do if people… Basically explaining the kind of limitations of the relationship between an AI and a human, that it doesn't retain things from the conversation. And so I think it will

just explain to you like, "Hey, I won't remember this conversation. Here's how I was trained. It's unlikely that I can have a certain kind of relationship with you, and it's important that you know that. It's important for your mental well-being that you don't think that I'm something that I'm not." And somehow I feel like this is one of the things where I'm like, "Ah, it feels like a thing that I always want to be true." I don't want models to be lying to people, because if people are going to have healthy relationships with anything, it's kind of… Yeah, I think that's easier if you always just know exactly what the thing is that you are relating to. It doesn't solve everything, but I think it helps quite a lot.

Amanda Askell
Anthropic may be the very company to develop a system that we definitively recognize as AGI, and you very well might be the person that talks to it, probably talks to it first. What would the conversation contain? What would be your first question?

Lex Fridman
Well, it depends partly on the capability level of the model. If you have something that is capable in the same way that an extremely capable human is, I imagine myself interacting with it the same way that I do with an extremely capable human, with the one difference that I'm probably going to be trying to probe and understand its behaviors. But in many ways, I'm like, I can then just have useful conversations with it. So, if I'm working on something as part of my research, I can just be like, "Oh." Which I already find myself starting to do. If I'm like, "Oh, I feel like there's this thing in virtue ethics. I can't quite remember the term," I'll use the model for things like that. And so I can imagine that being more and more the case where you're just basically interacting with it much more like you would an incredibly smart colleague and using it for the kinds of work that you want to do as if you just had a collaborator who was like… Or the slightly horrifying thing about AI is as soon as you have one collaborator, you have 1,000 collaborators if you can manage them enough.

Amanda Askell
But what if it's two times the smartest human on Earth on that particular discipline?

Lex Fridman
Yeah.

Amanda Askell
I guess you're really good at probing Claude in a way that pushes its limits, understanding where the limits are.

Lex Fridman
Yep.

Amanda Askell
So, I guess what would be a question you would ask to be like, "Yeah, this is AGI"?

Lex Fridman
That's really hard because it feels like it has to just be a series of questions. If there was just one question, you can train anything to answer one question extremely well. In fact, you can probably train it to answer 20 questions extremely well.

Amanda Askell
How long would you need to be locked in a room with an AGI to know this thing is AGI?

Lex Fridman
It's a hard question because part of me is like, "All of this just feels continuous." If you put me in a room for five minutes, I just have high error bars. And then it's just like, maybe it's both the probability increases and the error bar decreases. I think things that I can actually probe the edge of human knowledge of. So, I think this with philosophy a little bit. Sometimes when I ask the models philosophy questions, I am like, "This is a question that I think no one has ever asked." It's maybe right at the edge of some literature that I know. And the models, when they struggle with that, when they struggle to come up with a novel… I'm like, "I know that there's a novel argument here because I've just thought of it myself." So, maybe that's the thing where I'm like, "I've thought of a cool novel argument in this niche area, and I'm going to just probe you to see if you can come up with it and how much prompting it takes to get you to come up with it." And I think for some of these really right at the edge of human knowledge questions, I'm like, "You could not in fact come up with the thing that I came up with." I think if I just took something like that where I know a lot about an area and I came up with a novel issue or a novel solution to a problem, and I gave it to a model, and it came up with that solution, that would be a pretty moving moment for me because I would be like, "This is a case where no human has ever…" And obviously, you see novel solutions all the time, especially to easier problems. I think people overestimate that novelty isn't like… It's completely different from anything that's ever happened. It's just like it can be a variant of things that have happened and still be novel. But I think, yeah, the more I were to see completely novel work from the models that that would be… And this is just going to feel iterative. It's one of those things where there's never… It's like, people, I think, want there to be a moment, and I'm like, "I don't know." I think that there might just never be a moment. It might just be that there's just this continuous ramping up.

Amanda Askell
I have a sense that there would be things that a model can say that convinces you this is very… I've talked to people who are truly wise, because you could just tell there's a lot of horsepower there, and if you 10X that… I don't know. I just feel like there's words you could say. Maybe ask it to generate a poem, and the poem it generates, you're like, "Yeah, okay. Whatever you did there, I don't think a human can do that."

Lex Fridman
I think it has to be something that I can verify is actually really good, though. That's why I think these questions that are where I'm like, "Oh, this is like..." Sometimes it's just like I'll come up with, say, a concrete counter example to an argument or something like that. It would be like if you're a mathematician, you had a novel proof, I think, and you just gave it the problem, and you saw it, and you're like, "This proof is genuinely novel. You actually have to do a lot of things to come up with this. I had to sit and think about it for months," or something. And then if you saw the model successfully do that, I think you would just be like, "I can verify that this is correct. It is a sign that you have generalized from your training. You didn't just see this somewhere because I just came up with it myself, and you were able to replicate that." That's the kind of thing where I'm like, for me, the more that models can do things like that, the more I would be like, "Oh, this is very real." Because then, I don't know, I can verify that that's extremely, extremely capable.

Amanda Askell
You've interacted with AI a lot. What do you think makes humans special?

Lex Fridman
Oh, good question.

Amanda Askell
Maybe in a way that the universe is much better off that we're in it, and that we should definitely survive and spread throughout the universe.

Lex Fridman
Yeah, it's interesting because I think people focus so much on intelligence, especially with models. Look, intelligence is important because of what it does. It's very useful. It does a lot of things in the world. And I'm like, you can imagine a world where height or strength would have played this role, and it's just a trait like that. I'm like, it's not intrinsically valuable. It's valuable because of what it does, I think, for the most part. I mean, personally, I'm just like, I think humans and life in general is extremely magical. I don't know. Not everyone agrees with this. I'm flagging. But we have this whole universe, and there's all of these objects, there's beautiful stars and there's galaxies. Then, I don't know, I'm just like, on this planet there are these creatures that have this ability to observe that, and they are seeing it, they are experiencing it. And I'm just like, that, if you try to explain... I'm imagining trying to explain to, I don't know, someone. For some reason, they've never encountered the world, or science, or anything. And I think that everything, all of our physics and everything in the world, it's all extremely exciting. But then you say, "Oh, and plus there's this thing that it is to be a thing and observe in the world," and you see this inner cinema. And I think they would be like, "Hang on, wait, pause. You just said something that is kind of wild sounding." And so I'm like, we have this ability to experience the world. We feel pleasure, we feel suffering. We feel like a lot of complex things. Yeah. And maybe this is also why I think I also hear a lot

about animals, for example, because I think they probably share this with us. So, I think that the things that make humans special insofar as I care about humans is probably more like their ability to feel an experience than it is them having these functional, useful traits.

Amanda Askell
Yeah. To feel and experience the beauty in the world. Yeah. To look at the stars. I hope there's other alien civilizations out there, but if we're it, it's a pretty good thing.

Lex Fridman
And that they're having a good time.

Amanda Askell
A very good time watching us.

Lex Fridman
Yeah.

Amanda Askell
Well, thank you for this good time of a conversation and for the work you're doing and for helping make Claude a great conversational partner. And thank you for talking today.

Lex Fridman
Yeah, thanks for talking.

Amanda Askell
Thanks for listening to this conversation with Amanda Askell. And now, dear friends, here's Chris Olah. Can you describe this fascinating field of mechanistic interpretability, aka mech interp, the history of the field, and where it stands today?

Lex Fridman
I think one useful way to think about neural networks is that we don't program, we don't make them, we grow them. We have these neural network architectures that we design and we have these loss objectives that we create. And the neural network architecture, it's kind of like a scaffold that the circuits grow on. It starts off with some random things, and it grows, and it's almost like the objective that we train for is this light. And so we create the scaffold that it grows on, and we create the light that it grows towards. But the thing that we actually create, it's this almost biological entity or organism that we're studying. And so it's very, very different from any kind of regular software engineering because, at the end of the day, we end up with this artifact that can do all these amazing things. It can write essays and translate and understand images. It can do all these things that we have no idea how to directly create a computer program to do. And it can do that because we grew it. We didn't write it. We didn't create it. And so then that leaves open this question at the end, which is

what the hell is going on inside these systems? And that is, to me, a really deep and exciting question. It's a really exciting scientific question. To me, it is like the question that is just screaming out, it's calling out for us to go and answer it when we talk about neural networks. And I think it's also a very deep question for safety reasons.

Chris Olah
And mechanistic interpretability, I guess, is closer to maybe neurobiology?

Lex Fridman
Yeah, yeah, I think that's right. So, maybe to give an example of the kind of thing that has been done that I wouldn't consider to be mechanistic interpretability. There was, for a long time, a lot of work on saliency maps, where you would take an image and you'd try to say, "The model thinks this image is a dog. What part of the image made it think that it's a dog?" And that tells you maybe something about the model if you can come up with a principled version of that, but it doesn't really tell you what algorithms are running in the model, how is the model actually making that decision? Maybe it's telling you something about what was important to it, if you can make that method work, but it isn't telling you what are the algorithms that are running? How is it that the system's able to do this thing that no one knew how to do? And so I guess we started using the term mechanistic interpretability to try to draw that divide or to distinguish ourselves in the work that we were doing in some ways from some of these other things. And I think since then, it's become this sort of umbrella term for a pretty wide variety of work. But I'd say that the things that are kind of distinctive are, I think, A, this focus on, we really want to get at the mechanisms. We want to get at algorithms. If you think of neural networks as being like a computer program, then the weights are kind of like a binary computer program. And we'd like to reverse engineer those weights and figure out what algorithms are running. So okay, I think one way you might think of trying to understand a neural network is that it's kind of like we have this compiled computer program, and the weights of the neural network are the binary. And when the neural network runs, that's the activations. And our goal is ultimately to go and understand these weights. And so the project of mechanistic interpretability is to somehow figure out how do these weights correspond to algorithms? And in order to do that, you also have to understand the activations because the activations are like the memory. And if you imagine reverse engineering a computer program, and you have the binary instructions, in order to understand what a particular instruction means, you need to know what is stored in the memory that it's operating on. And so those two things are very intertwined. So, mechanistic interpretability tends to be interested in both of those things. Now, there's a lot of work that's interested in those things, especially there's all this work on probing, which you might see as part of being mechanistic interpretability, although, again, it's just a broad term, and not everyone who does that work would identify as doing mech interp. I think a thing that is maybe a little bit distinctive to the vibe of mech interp is I think people working in this space tend to think of neural networks as... Well, maybe one way to say it is the gradient descent is smarter than you. That gradient descent is actually really great. The

whole reason that we're understanding these models is because we didn't know how to write them in the first place. The gradient descent comes up with better solutions than us. And so I think that maybe another thing about mech interp is having almost a kind of humility, that we won't guess a priori what's going on inside the model. We have to have this sort of bottom up approach where we don't assume that we should look for a particular thing, and that will be there, and that's how it works. But instead, we look for the bottom up and discover what happens to exist in these models and study them that way.

Chris Olah
But the very fact that it's possible to do, and as you and others have shown over time, things like universality, that the wisdom of the gradient descent creates features and circuits, creates things universally across different kinds of networks that are useful, and that makes the whole field possible.

Lex Fridman
Yeah. So this, actually, is indeed a really remarkable and exciting thing, where it does seem like, at least to some extent, the same elements, the same features and circuits, form again and again. You can look at every vision model, and you'll find curve detectors, and you'll find high-low-frequency detectors. And in fact, there's some reason to think that the same things form across biological neural networks and artificial neural networks. So, a famous example is vision models in the early layers. They have Gabor filters, and Gabor filters are something that neuroscientists are interested in and have thought a lot about. We find curve detectors in these models. Curve detectors are also found in monkeys. We discover these high-low-frequency detectors, and then some follow-up work went and discovered them in rats or mice. So, they were found first in artificial neural networks and then found in biological neural networks. There's this really famous result on grandmother neurons or the Halle Berry neuron from Quiroga et al. And we found very similar things in vision models, where this is while I was still at OpenAI, and I was looking at their clip model, and you find these neurons that respond to the same entities in images. And also, to give a concrete example there, we found that there was a Donald Trump neuron. For some reason, I guess everyone likes to talk about Donald Trump. And Donald Trump was very prominent, was a very hot topic at that time. So, every neural network we looked at, we would find a dedicated neuron for Donald Trump. That was the only person who had always had a dedicated neuron. Sometimes you'd have an Obama neuron, sometimes you'd have a Clinton neuron, but Trump always had a dedicated neuron. So, it responds to pictures of his face and the word Trump, all of these things, right? And so it's not responding to a particular example, or it's not just responding to his face, it's abstracting over this general concept. So in any case, that's very similar to these Quiroga et al results. So, this evidence that this phenomenon of universality, the same things form across both artificial and natural neural networks, that's a pretty amazing thing if that's true. Well, I think the thing that suggests is that gradient descent is finding the right ways to cut things apart, in some sense, that many systems converge on and many different neural networks architectures converge on. Now

there's some set of abstractions that are a very natural way to cut apart the problem and that a lot of systems are going to converge on. I don't know anything about neuroscience. This is just my wild speculation from what we've seen.

Chris Olah
Yeah. That would be beautiful if it's sort of agnostic to the medium of the model that's used to form the representation.

Lex Fridman
Yeah, yeah. And it's kind of a wild speculation-based… We only have a few data points that's just this, but it does seem like there's some sense in which the same things form again and again both certainly in natural neural networks and also artificially, or in biology.

Chris Olah
And the intuition behind that would be that in order to be useful in understanding the real world, you need all the same kind of stuff.

Lex Fridman
Yeah. Well, if we pick, I don't know, the idea of a dog, right? There's some sense in which the idea of a dog is like a natural category in the universe, or something like this. There's some reason. It's not just a weird quirk of how humans think about the world that we have this concept of a dog. Or if you have the idea of a line. Look around us. There are lines. It's the simplest way to understand this room, in some sense, is to have the idea of a line. And so I think that that would be my instinct for why this happens.

Chris Olah
Yeah. You need a curved line to understand a circle, and you need all those shapes to understand bigger things. And it's a hierarchy of concepts that are formed. Yeah.

Lex Fridman
And maybe there are ways to go and describe images without reference to those things, right? But they're not the simplest way, or the most economical way, or something like this. And so systems converge to these strategies would be my wild hypothesis.

Chris Olah
Can you talk through some of the building blocks that we've been referencing of features and circuits? So, I think you first described them in a 2020 paper, Zoom In: An Introduction to Circuits.

Lex Fridman
Absolutely. So, maybe I'll start by just describing some phenomena, and then we can build to the idea of features and circuits.

Chris Olah
Wonderful.

Lex Fridman
So, if you spent quite a few years, maybe five years, to some extent, with other things, studying this one particular model, Inception V1, which is this one vision model… It was state-of-the-art in 2015, and very much not state-of-the-art anymore. And it has maybe about 10,000 neurons in it. I spent a lot of time looking at the 10,000 neurons, odd neurons of Inception V1. One of the interesting things is there are lots of neurons that don't have some obvious interpretable meaning, but there's a lot of neurons in Inception V1 that do have really clean interpretable meanings. So, you find neurons that just really do seem to detect curves, and you find neurons that really do seem to detect cars, and car wheels, and car windows, and floppy ears of dogs, and dogs with long snouts facing to the right, and dogs with long snouts facing to the left, and different kinds of fur. And there's this whole beautiful edge detectors, line detectors, color contrast detectors, these beautiful things we call high-low-frequency detectors. I think looking at it, I sort of felt like a biologist. You're looking at this sort of new world of proteins, and you're discovering all these different proteins that interact. So, one way you could try to understand these models is in terms of neurons. You could try to be like, "Oh, there's a dog detecting neuron, and here's a car detecting neuron." And it turns out you can actually ask how those connect together. So, you can go say, "Oh, I have this car detecting neuron. How was it built?" And it turns out, in the previous layer, it's connected really strongly to a window detector, and a wheel detector, and a car body detector. And it looks for the window above the car, and the wheels below, and the car chrome in the middle, sort of everywhere, but especially on the lower part. And that's sort of a recipe for a car, right? Earlier, we said the thing we wanted from mech interp was to get algorithms to go and get, ask, "What is the algorithm that runs?" Well, here we're just looking at the weights of the neural network and we're reading off this recipe for detecting cars. It's a very simple, crude recipe, but it's there. And so we call that a circuit, this connection. Well, okay, so the problem is that not all of the neurons are interpretable. And there's reason to think, we can get into this more later, that there's this superposition hypothesis, there's reason to think that sometimes the right unit to analyze things is combinations of neurons. So, sometimes it's not that there's a single neuron that represents, say, a car, but it actually turns out after you detect the car, the model hides a little bit of the car in the following layer, in a bunch of dog detectors. Why is it doing that? Well, maybe it just doesn't want to do that much work on cars at that point, and it's storing it away to go and… So, it turns out, then, that this sort of subtle pattern of… There's all these neurons that you think are dog detectors, and maybe they're primarily that, but they all a little bit contribute to representing a car in that next layer. Okay? So, now we can't really think… There might still be something, I don't know, you could call it a car concept or something, but it no longer corresponds to a neuron. So, we need some term for these kind of neuron-like entities, these things that we would have liked the neurons to be, these

idealized neurons. The things that are the nice neurons, but also maybe there's more of them somehow hidden. And we call those features.

Chris Olah
And then what are circuits?

Lex Fridman
So, circuits are these connections of features, right? So, when we have the car detector and it's connected to a window detector and a wheel detector, and it looks for the wheels below and the windows on top, that's a circuit. So, circuits are just collections of features connected by weights, and they implement algorithms. So, they tell us how are features used, how are they built, how do they connect together? So, maybe it's worth trying to pin down what really is the core hypothesis here? And I think the core hypothesis is something we call the linear representation hypothesis. So, if we think about the car detector, the more it fires, the more we think of that as meaning, "Oh, the model is more and more confident that a car is present." Or if it's some combination of neurons that represent a car, the more that combination fires, the more we think the model thinks there's a car present. This doesn't have to be the case, right? You could imagine something where you have this car detector neuron and you think, "Ah, if it fires between one and two, that means one thing, but it means something totally different if it's between three and four." That would be a nonlinear representation. And in principle, models could do that. I think it's sort of inefficient for them to do. If you try to think about how you'd implement computation like that, it's kind of an annoying thing to do. But in principle, models can do that. So, one way to think about the features and circuits sort of framework for thinking about things is that we're thinking about things as being linear. We're thinking about that if a neuron or a combination of neurons fires more, that means more of a particular thing being detected. And then that gives weight, a very clean interpretation as these edges between these entities that these features, and that that edge then has a meaning. So that's, in some ways, the core thing. It's like we can talk about this outside the context of neurons. Are you familiar with the Word2Vec results?

Chris Olah
Mm- hmm.

Lex Fridman
You have king – man + woman = queen. Well, the reason you can do that kind of arithmetic is because you have a linear representation.

Chris Olah
Can you actually explain that representation a little bit? So first off, the feature is a direction of activation.

Lex Fridman
Yeah, exactly.

Chris Olah
You can do it that way. Can you do the – men + women, that, the Word2Vec stuff? Can you explain what that is, that work?

Lex Fridman
Yeah. So, there's this very-

Chris Olah
It's such a simple, clean explanation of what we're talking about.

Lex Fridman
Exactly. Yeah. So, there's this very famous result, Word2Vec, by Tomas Mikolov et al, and there's been tons of follow-up work exploring this. So, sometimes we create these word embeddings where we map every word to a vector. I mean, that in itself, by the way, is kind of a crazy thing if you haven't thought about it before, right?

Chris Olah
Mm-hmm.

Lex Fridman
If you just learned about vectors in physics class, and I'm like, "Oh, I'm going to actually turn every word in the dictionary into a vector," that's kind of a crazy idea. Okay. But you could imagine all kinds of ways in which you might map words to vectors. But it seems like when we train neural networks, they like to go and map words to vectors such that there's sort of linear structure in a particular sense, which is that directions have meaning. So, for instance, there will be some direction that seems to sort of correspond to gender, and male words will be far in one direction, and female words will be in another direction. And the linear representation hypothesis is, you could think of it roughly as saying that that's actually the fundamental thing that's going on, that everything is just different directions have meanings, and adding different direction vectors together can represent concepts. And the Mikolov paper took that idea seriously, and one consequence of it is that you can do this game of playing arithmetic with words. So, you can do king and you can subtract off the word man and add the word woman. And so you're sort of going and trying to switch the gender. And indeed, if you do that, the result will sort of be close to the word queen. And you can do other things like you can do sushi – Japan + Italy and get pizza, or different things like this, right? So this is, in some sense, the core of the linear representation hypothesis. You can describe it just as a purely abstract thing about vector spaces. You can describe it as a statement about the activations of neurons, but it's really about this property of directions having meaning. And in some ways, it's even a little subtler than… It's really, I

think, mostly about this property of being able to add things together, that you can independently modify, say gender and royalty, or cuisine type, or country, and the concept of food by adding them.

Chris Olah

Do you think the linear hypothesis holds -

Lex Fridman

Yes.

Chris Olah

… that carries scales?

Lex Fridman

So far, I think everything I have seen is consistent with this hypothesis, and it doesn't have to be that way, right? You can write down neural networks where you write weights such that they don't have linear representations, where the right way to understand them is not in terms of linear representations. But I think every natural neural network I've seen has this property. There's been one paper recently that there's been some sort of pushing around the edge. So, I think there's been some work recently studying multidimensional features where rather than a single direction, it's more like a manifold of directions. This, to me, still seems like a linear representation. And then there's been some other papers suggesting that maybe in very small models you get non-linear representations. I think that the jury's still out on that. But I think everything that we've seen so far has been consistent with the linear representation hypothesis, and that's wild. It doesn't have to be that way. And yet I think that there's a lot of evidence that certainly at least this is very, very widespread, and so far the evidence is consistent with that. And I think one thing you might say is you might say, "Well, Christopher, that's a lot to go and to ride on. If we don't know for sure this is true, and you're investing it in neural networks as though it is true, isn't that dangerous?" But I think, actually, there's a virtue in taking hypotheses seriously and pushing them as far as they can go. So, it might be that someday we discover something that isn't consistent with a linear representation hypothesis, but science is full of hypotheses and theories that were wrong, and we learned a lot by working under them as an assumption and then going and pushing them as far as we can. I guess this is the heart of what Kuhn would call normal science. I don't know. If you want, we can talk a lot about-

Chris Olah

Kuhn.

Lex Fridman

… philosophy of science and -

Chris Olah
That leads to the paradigm shift. So yeah, I love it, taking the hypothesis seriously, and take it to a natural conclusion.

Lex Fridman
Yeah.

Chris Olah
Same with the scaling hypothesis. Same -

Lex Fridman
Exactly. Exactly. And -

Chris Olah
I love it.

Lex Fridman
One of my colleagues, Tom Henighan, who is a former physicist, made this really nice analogy to me of caloric theory where once upon a time, we thought that heat was actually this thing called caloric. And the reason hot objects would warm up cool objects is the caloric is flowing through them. And because we're so used to thinking about heat in terms of the modern theory, that seems kind of silly. But it's actually very hard to construct an experiment that disproves the caloric hypothesis. And you can actually do a lot of really useful work believing in caloric. For example, it turns out that the original combustion engines were developed by people who believed in the caloric theory. So, I think there's a virtue in taking hypotheses seriously even when they might be wrong.

Chris Olah
Yeah, there's a deep philosophical truth to that. That's kind of how I feel about space travel, like colonizing Mars. There's a lot of people that criticize that. I think if you just assume we have to colonize Mars in order to have a backup for human civilization, even if that's not true, that's going to produce some interesting engineering and even scientific breakthroughs, I think.

Lex Fridman
Yeah. Actually, this is another thing that I think is really interesting. So, there's a way in which I think it can be really useful for society to have people almost irrationally dedicated to investigating particular hypotheses because, well, it takes a lot to maintain scientific morale and really push on something when most scientific hypotheses end up being wrong. A lot of science doesn't work out, and yet it's very useful to... There's a joke about Geoff Hinton, which is that Geoff Hinton has discovered how the brain works every year for the

last 50 years. But I say that with really deep respect because, in fact, actually, that led to him doing some really great work.

Chris Olah

Yeah, he won the Nobel Prize now. Who's laughing now?

Lex Fridman

Exactly. Exactly. Exactly. I think one wants to be able to pop up and recognize the appropriate level of confidence. But I think there's also a lot of value in just being like, "I'm going to essentially assume, I'm going to condition on this problem being possible or this being broadly the right approach. And I'm just going to go and assume that for a while and go and work within that, and push really hard on it." And if society has lots of people doing that for different things, that's actually really useful in terms of going and- … things that's actually really useful in terms of going and either really ruling things out. We can be like, "Well, that didn't work and we know that somebody tried hard." Or going and getting to something that does teach us something about the world.

Chris Olah

So another interesting hypothesis is the super superposition hypothesis. Can you describe what superposition is?

Lex Fridman

Yeah. So earlier we were talking about word defect, right? And we were talking about how maybe you have one direction that corresponds to gender and maybe another that corresponds to royalty and another one that corresponds to Italy and another one that corresponds to food and all of these things. Well, oftentimes maybe these word embeddings, they might be 500 dimensions, a thousand dimensions. And so if you believe that all of those directions were orthogonal, then you could only have 500 concepts. And I love pizza. But if I was going to go and give the 500 most important concepts in the English language, probably Italy wouldn't be… it's not obvious, at least that Italy would be one of them, right? Because you have to have things like plural and singular and verb and noun and adjective. And there's a lot of things we have to get to before we get to Italy and Japan, and there's a lot of countries in the world. And so how might it be that models could simultaneously have the linear representation hypothesis be true and also represent more things than they have directions? So what does that mean? Well, okay, so if linear representation hypothesis is true, something interesting has to be going on. Now, I'll tell you one more interesting thing before we go, and we do that, which is earlier we were talking about all these polysemantic neurons, these neurons that when we were looking at inception V1, these nice neurons that the car detector and the curve detector and so on that respond to lots of very coherent things. But it's lots of neurons that respond to a bunch of unrelated things. And that's also an interesting phenomenon. And it turns out as well that even these neurons that are really, really clean, if you look at the weak activations, so if you look at the activations where it's activating 5% of the maximum activation, it's really not the core thing that it's expecting. So if you look at a curve detector for instance, and you look at the places where it's 5% active, you could interpret it just as noise or it could be that it's doing something else there. Okay? So how could that be? Well, there's this amazing thing in mathematics called compressed sensing, and it's actually this very surprising fact where if you have a high dimensional space and you project it into a low dimensional space, ordinarily you can't go and sort of un-projected and get back your high dimensional vector, you threw information away. This is like you can't invert a rectangular matrix. You can only invert square matrices. But it turns out that that's actually not quite true. If I tell you that the high-dimensional vector was sparse, so it's mostly zeros, then it turns out that you can often go and find back the high-dimensional vector with very high probability. So that's a surprising fact, right? It says that you can have this high-dimensional vector space, and as long as things are sparse, you can project it down, you can have a lower-dimensional projection of it, and that works. So the superstition hypothesis is saying that that's what's going on in neural networks, for instance, that's what's going on in word embeddings. The word embeddings are able to simultaneously have directions be the meaningful thing, and by exploiting the fact that they're operating on a fairly high-dimensional space, they're actually… and the fact that these concepts are sparse, you usually aren't talking about Japan and Italy at the same time. Most of those concepts, in most instances, Japan and Italy are both zero. They're not present at all. And if that's true, then you can go and have it be the case that you can have many more of these sort of directions that are meaningful, these features than you have dimensions. And similarly, when we're talking about neurons, you can have many more concepts than you have neurons. So that's at a high level, the superstition hypothesis. Now it has this even wilder implication, which is to go and say that neural networks, it may not just be the case that the representations are like this, but the computation may also be like this. The connections between all of them. And so in some sense, neural networks may be shadows of much larger sparser neural networks. And what we see are these projections. And the strongest version of superstition hypothesis would be to take that really seriously and sort of say there actually is in some sense this upstairs model where the neurons are really sparse and all interpleural, and the weights between them are these really sparse circuits. And that's what we're studying. And the thing that we're observing is the shadow of evidence. We need to find the original object.

Chris Olah

And the process of learning is trying to construct a compression of the upstairs model that doesn't lose too much information in the projection.

Lex Fridman

Yeah, it's finding how to fit it efficiently or something like this. The gradient descent is doing this and in fact, so this sort of says that gradient descent, it could just represent a dense neural network, but it sort of says that gradient descent is implicitly searching over the space of extremely sparse models that could be projected into this low-dimensional space.

And this large body of work of people going and trying to study sparse neural networks where you go and you have… you could design neural networks where the edges are sparse and the activations are sparse. And my sense is that work has generally, it feels very principled, it makes so much sense, and yet that work hasn't really panned out that well, is my impression broadly. And I think that a potential answer for that is that actually the neural network is already sparse in some sense. You were trying to go and do this. Gradient descent was actually behind the scenes going and searching more efficiently than you could through the space of sparse models and going and learning whatever sparse model was most efficient. And then figuring out how to fold it down nicely to go and run conveniently on your GPU, which does as nice dense matrix multiplies. And that you just can't beat that.

Chris Olah
How many concepts do you think can be shoved into a neural network?

Lex Fridman
Depends on how sparse they are. So there's probably an upper bound from the number of parameters because you still have to have print weights that go and connect them together. So that's one upper bound. There are in fact all these lovely results from compressed sensing and the Johnson-Lindenstrauss lemma and things like this that they basically tell you that if you have a vector space and you want to have almost orthogonal vectors, which is sort of probably the thing that you want here. So you're going to say, "Well, I'm going to give up on having my concepts, my features be strictly orthogonal, but I'd like them to not interfere that much. I'm going to have to ask them to be almost orthogonal." Then this would say that it's actually for, once you set a threshold for what you're willing to accept in terms of how much cosine similarity there is, that's actually exponential in the number of neurons that you have. So at some point, that's not going to even be the limiting factor, but there's some beautiful results there. And in fact, it's probably even better than that in some sense because that's sort of for saying that any random set of features could be active. But in fact the features have sort of a correlational structure where some features are more likely to co-occur and other ones are less likely to co-occur. And so neural networks, my guest would be, could do very well in terms of going and packing things to the point that's probably not the limiting factor.

Chris Olah
How does the problem of polysemanticity enter the picture here?

Lex Fridman
Polysemanticity is this phenomenon we observe where you look at many neurons and the neuron doesn't just sort of represent one concept, it's not a clean feature. It responds to a bunch of unrelated things. And superstition you can think of as being a hypothesis that explains the observation of polysemanticity. So polysemanticity is this observed

phenomenon and superstition is a hypothesis that would explain it along with some other things.

Chris Olah
So that makes Mechinterp more difficult.

Lex Fridman
Right. So if you're trying to understand things in terms of individual neurons and you have polysemantic neurons, you're in an awful lot of trouble. The easiest answer is like, "Okay, well you're looking at the neurons, you're trying to understand them. This one responds for a lot of things. It doesn't have a nice meaning. Okay, that's bad." Another thing you could ask is ultimately we want to understand the weights. And if you have two polysemantic neurons and each one responds to three things and then the other neuron responds to three things and you have a wait between them, what does that mean? Does it mean that all three, there's these nine interactions going on? It's a very weird thing, but there's also a deeper reason, which is related to the fact that neural networks operate on really high dimensional spaces. So I said that our goal was to understand neural networks and understand the mechanisms. And one thing you might say is, "Well, it's just a mathematical function. Why not just look at it, right?" One of the earliest projects I did studied these neural networks that mapped two-dimensional spaces to two-dimensional spaces, and you can sort of interpret them in this beautiful way is like bending manifolds. Why can't we do that? Well, as you have a higher dimensional space, the volume of that space in some sense is exponential in the number of inputs you have. And so you can't just go and visualize it. So we somehow need to break that apart. We need to somehow break that exponential space into a bunch of things, some non-exponential number of things that we can reason about independently. And the independence is crucial because it's the independence that allows you to not have to think about all the exponential combinations of things. And things being monosomatic, things only having one meaning, things having a meaning, that is the key thing that allows you to think about them independently. And so I think if you want the deepest reason why we want to have interpretable monosomatic features, I think that's really the deep reason.

Chris Olah
And so the goal here as your recent work has been aiming at is how do we extract the monosomatic features from a neural net that has polysemantic features and all this mess.

Lex Fridman
Yes, we observe these polysemantic neurons, we hypothesize that's what's going on is superposition. And if superposition is what's going on, there's actually a sort of well-established technique that is sort of the principled thing to do, which is dictionary learning. And it turns out if you do dictionary learning in particular, if you do sort of a nice efficient way that in some sense sort of nicely regularizes that as well called a sparse auto encoder. If you train a sparse auto encoder, these beautiful interpretable features start to

just fall out where there weren't any beforehand. So that's not a thing that you would necessarily predict, but it turns out that works very, very well. To me, that seems like some non-trivial validation of linear representations and superposition.

Chris Olah
So with dictionary learning, you're not looking for particular kind of categories. You don't know what they are, they just emerge.

Lex Fridman
Exactly. And this gets back to our earlier point when we're not making assumptions. Gradient descent is smarter than us, so we're not making assumptions about what's there. I mean, one certainly could do that, right? One could assume that there's a PHP feature and go and search for it, but we're not doing that. We're saying we don't know what's going to be there. Instead, we're just going to go and let the sparse auto encoder discover the things that are there.

Chris Olah
So can you talk toward monosematicity paper from October last year? I heard a lot of nice breakthrough results.

Lex Fridman
That's very kind of you to describe it that way. Yeah, I mean, this was our first real success using sparse autoencoders. So we took a one-layer model, and it turns out if you go and you do dictionary learning on it, you find all these really nice interpretable features. So the Arabic feature, the Hebrew feature, the Base64 features were some examples that we studied in a lot of depth and really showed that they were what we thought they were. Turns out if you train a model twice as well and train two different models and do dictionary learning, you find analogous features in both of them. So that's fun. You find all kinds of different features. So that was really just showing that this works. And I should mention that there was this Cunningham and all that had very similar results around the same time.

Chris Olah
There's something fun about doing these kinds of small scale experiments and finding that it's actually working.

Lex Fridman
Yeah, well, and that there's so much structure here. So maybe stepping back, for a while I thought that maybe all this mechanistic interpolate work, the end result was going to be that I would have an explanation for why it was sort of very hard and not going to be tractable. We'd be like, "Well, there's this problem with supersession and it turns out supersession is really hard and we're kind of screwed, but that's not what happened. In fact, a very natural simple technique just works. And so then that's actually a very good situation.

I think this is a sort of hard research problem and it's got a lot of research risk and it might still very well fail, but I think that some very significant amount of research risk was put behind us when that started to work.

Chris Olah
Can you describe what kind of features can be extracted in this way?

Lex Fridman
Well, so it depends on the model that you're studying. So the larger the model, the more sophisticated they're going to be. And we'll probably talk about follow up work in a minute. But in these one layer models, so some very common things I think were languages, both programming languages and natural languages. There were a lot of features that were specific words in specific contexts, so the. And I think really the way to think about this is that the is likely about to be followed by a noun. So you could think of this as the feature, but you could also think of this as protecting a specific noun feature. And there would be these features that would fire for the in the context of say, a legal document or a mathematical document or something like this. And so maybe in the context of math, you're like the, and then predict vector or matrix, all these mathematical words, whereas in other contexts you would predict other things, that was common.

Chris Olah
And basically we need clever humans to assign labels to what we're seeing.

Lex Fridman
Yes. So the only thing this is doing is that sort of unfolding things for you. So if everything was sort of folded over top of it, serialization folded everything on top of itself and you can't really see it, this is unfolding it. But now you still have a very complex thing to try to understand. So then you have to do a bunch of work understanding what these are, and some are really subtle. There's some really cool things even in this one layer model about Unicode, where of course some languages are in Unicode, and the tokenizer won't necessarily have a dedicated token for every Unicode character. So instead, what you'll have is you'll have these patterns of alternating token or alternating tokens that each represent half of a Unicode character. And you have a different feature that goes and activates on the opposing ones to be like, "Okay, I just finished a character, go and predict next prefix. Then okay, I'm on the prefix, predict a reasonable suffix." And you have to alternate back and forth. So these swap layer models are really interesting. And I mean there's another thing that you might think, "Okay, there would just be one Base64 feature, but it turns out there's actually a bunch of Base64 features because you can have English text encoded as Base64, and that has a very different distribution of Base64 tokens than regular. And there's some things about tokenization as well that it can exploit. And I don't know, there's all kinds of fun stuff.

Chris Olah
How difficult is the task of assigning labels to what's going on? Can this be automated by AI?

Lex Fridman
Well, I think it depends on the feature, and it also depends on how much you trust your AI. So there's a lot of work doing automated interoperability. I think that's a really exciting direction, and we do a fair amount of automated interoperability and have Claude go and label our features.

Chris Olah
Is there some fun moments where it's totally right or it's totally wrong?

Lex Fridman
Yeah, well, I think it's very common that it says something very general, which is true in some sense, but not really picking up on the specific of what's going on. So I think that's a pretty common situation. You don't know that I have a particularly amusing one.

Chris Olah
That's interesting. That little gap between it is true, but it doesn't quite get to the deep nuance of a thing. That's a general challenge, it's already an incredible caution that can say a true thing, but it's missing the depth sometimes. And in this context, it's like the ARC challenge, the sort of IQ type of tests. It feels like figuring out what a feature represents is a little puzzle you have to solve.

Lex Fridman
Yeah. And I think that sometimes they're easier and sometimes they're harder as well. Yeah, I think that's tricky. There's another thing which I don't know, maybe in some ways this is my aesthetic coming in, but I'll try to give you a rationalization. I'm actually a little suspicious of automated interoperability, and I think that partly just that I want humans to understand neural networks. And if the neural network is understanding it for me, I don't quite like that, but I do have a bit of… In some ways, I'm sort of like the mathematicians who are like, "If there's a computer automated proof, it doesn't count." They won't understand it. But I do also think that there is this kind of reflections on trusting trust type issue where there's this famous talk about when you're writing a computer program, you have to trust your compiler. And if there was malware in your compiler, then it could go and inject malware into the next compiler and you'd be kind of in trouble, right? Well, if you're using neural networks to go and verify that your neural networks are safe, the hypothesis that you're trusting for is like, "Okay, well the neural network maybe isn't safe and you have to worry about is there some way that it could be screwing with you? I think that's not a big concern now, but I do wonder in the long run, if we have to use really powerful AI systems to go and audit our AI systems, is that actually something we can trust? But maybe I'm just rationalizing because I just want us to have to get to a point where humans understand everything.

Chris Olah
Yeah, I mean that's hilarious, especially as we talk about AI safety and looking for features that would be relevant to AI safety, like deception and so on. So let's talk about the Scaling Monosematicity paper in May 2024. Okay. So what did it take to scale this, to apply to Claude 3 Sonnet?

Lex Fridman
Well, a lot of GPUs.

Chris Olah
A lot more GPUs. Got it.

Lex Fridman
But one of my teammates, Tom Henighan was involved in the original scaling laws work, and something that he was sort of interested in from very early on is are there scaling laws for interoperability? And so something he immediately did when this work started to succeed, and we started to have sparse autoencoders work, was he became very interested in what are the scaling laws for making sparse autoencoders larger and how does that relate to making the base model larger? And so it turns out this works really well and you can use it to sort of project, if you train a sparse autoencoder of a given size, how many tokens should you train on and so on. This was actually a very big help to us in scaling up this work, and made it a lot easier for us to go and train really large sparse autoencoders where it's not training the big models, but it's starting to get to a point where it's actually expensive to go and train the really big ones.

Chris Olah
I mean, you have to do all this stuff of splitting it across large CPUs-

Lex Fridman
Oh, yeah. No, I mean there's a huge engineering challenge here too, right? Yeah. So there's a scientific question of how do you scale things effectively? And then there's an enormous amount of engineering to go and scale this up. You have to chart it, you have to think very carefully about a lot of things. I'm lucky to work with a bunch of great engineers because I am definitely not a great engineer.

Chris Olah
And the infrastructure especially. Yeah, for sure. So it turns out TLDR, it worked.

Lex Fridman
It worked. Yeah. And I think this is important because you could have imagined a world where you set after towards monospecificity. Chris, this is great. It works on a one-layer model, but one-layer models are really idiosyncratic. Maybe that's just something, maybe

the linear representation hypothesis and superposition hypothesis is the right way to understand a one-layer model, but it's not the right way to understand larger models. So I think, I mean, first of all, the Cunningham and all paper sort of cut through that a little bit and sort of suggested that this wasn't the case. But Scaling Monospecificity sort of I think was significant evidence that even for very large models, and we did it on Claude 3 Sonnet, which at that point was one of our production models. Even these models seemed to be substantially explained, at least by linear features. And doing dictionary learning on them works, and as you learn more features, you go and you explain more and more. So that's, I think, quite a promising sign. And you find now really fascinating abstract features, and the features are also multimodal. They respond to images and texts for the same concept, which is fun.

Chris Olah
Yeah. Can you explain that? I mean, backdoor, there's just a lot of examples that you can-

Lex Fridman
Yeah. So maybe let's start with that. One example to start, which is we found some features around security vulnerabilities and backdoorsing code. So turns out those are actually two different features. So there's a security vulnerability feature, and if you force it active, Claude it will start to go and write security vulnerabilities like buffer overflows into code. And also fires for all kinds of things, some of the top data set examples where things like dash dash, disable SSL or something like this, which are sort of obviously really insecure.

Chris Olah
So at this point, maybe it's just because the examples are presented that way, it's kind of surface a little bit more obvious examples. I guess the idea is that down the line it might be able to detect more nuance like deception or bugs or that kind of stuff.

Lex Fridman
Yeah. Well, maybe I want to distinguish two things. So one is the complexity of the feature or the concept, right? And the other is the nuance of how subtle the examples we're looking at, right?. So when we show the top data set examples, those are the most extreme examples that cause that feature to activate. And so it doesn't mean that it doesn't fire for more subtle things. So that insecure code feature, the stuff that it fires most strongly for are these really obvious disable the security type things, but it also fires for buffer overflows and more subtle security vulnerabilities in code. These features are all multimodal. You could ask it like, "What images activate this feature?" And it turns out that the security vulnerability feature activates for images of people clicking on Chrome to go past this website, the SSL certificate might be wrong or something like this. Another thing that's very entertaining is there's backdoors in code feature, like you activate it, it goes and Claude writes a backdoor that will go and dump your data to port or something. But you can ask, "Okay, what images activate the backdoor feature?" It was devices with hidden cameras in

them. So there's a whole apparently genre of people going and selling devices that look innocuous that have hidden cameras, and they have ads that has this hidden camera in it? And I guess that is the physical version of a backdoor. And so it sort of shows you how abstract these concepts are, and I just thought that was… I'm sort of sad that there's a whole market of people selling devices like that, but I was kind of delighted that that was the thing that it came up with as the top image examples for the feature.

Chris Olah
Yeah, it's nice. It's multimodal. It's multi almost context. It's broad, strong definition of a singular concept. It's nice.

Lex Fridman
Yeah.

Chris Olah
To me, one of the really interesting features, especially for AI safety, is deception and lying. And the possibility that these kinds of methods could detect lying in a model, especially get smarter and smarter and smarter. Presumably that's a big threat over super intelligent model that it can deceive the people operating it as to its intentions or any of that kind of stuff. So what have you learned from detecting lying inside models?

Lex Fridman
Yeah, so I think we're in some ways in early days for that, we find quite a few features related to deception and lying. There's one feature where it fires for people lying and being deceptive, and you force it active and Claude starts lying to you. So we have a deception feature. I mean, there's all kinds of other features about withholding information and not answering questions, features about power seeking and coups and stuff like that. So there's a lot of features that are kind of related to spooky things, and if you force them active Claude will behave in ways that are… they're not the kinds of behaviors you want.

Chris Olah
What are possible next exciting directions to you in the space of Mechinterp?

Lex Fridman
Well, there's a lot of things. So for one thing, I would really like to get to a point where we have shortcuts where we can really understand not just the features, but then use that to understand the computation of models. That relief for me is the ultimate goal of this. And there's been some work, we put out a few things. There's a paper from Sam Marks that does some stuff like this, and there's been, I'd say some work around the edges here. But I think there's a lot more to do, and I think that will be a very exciting thing that's related to a challenge we call interference weights. Where due to superstition, if you just sort of naively look at what features are connected together, there may be some weights that don't exist in

the upstairs model, but are just sort of artifacts of superstition. So that's a technical challenge Related to that, I think another exciting direction is just you might think of sparse autoencoders as being kind of like a telescope. They allow us to look out and see all these features that are out there, and as we build better and better sparse autoencoders, we better and better at dictionary learning, we see more and more stars. And we zoom in on smaller and smaller stars. There's a lot of evidence that we're only still seeing a very small fraction of the stars. There's a lot of matter in our neural network universe that we can't observe yet. And it may be that we'll never be able to have fine enough instruments to observe it, and maybe some of it just isn't possible, isn't computationally tractable to observe. So it's sort of a kind of dark matter in not in maybe the sense of modern astronomy of early astronomy when we didn't know what this unexplained matter is. And so I think a lot about that dark matter and whether we'll ever observe it and what that means for safety if we can't observe it, if some significant fraction of neural networks are not accessible to us. Another question that I think a lot about is at the end of the day, mechanistic interpolation is this very microscopic approach to interpolation. It's trying to understand things in a very fine-grained way, but a lot of the questions we care about are very macroscopic. We care about these questions about neural network behavior, and I think that's the thing that I care most about. But there's lots of other sort of larger-scale questions you might care about. And the nice thing about having a very microscopic approach is it's maybe easier to ask, is this true? But the downside is its much further from the things we care about. And so we now have this ladder to climb. And I think there's a question of will we be able to find, are there larger-scale abstractions that we can use to understand neural networks that can we get up from this very microscopic approach?

Chris Olah
Yeah. You've written about this as kind of organs question.

Lex Fridman
Yeah, exactly.

Chris Olah
If we think of interpretability as a kind of anatomy of neural networks, most of the circus threads involve studying tiny little veins looking at the small scale and individual neurons and how they connect. However, there are many natural questions that the small-scale approach doesn't address. In contrast, the most prominent abstractions and biological anatomy involve larger-scale structures like individual organs, like the heart or entire organ systems like the respiratory system. And so we wonder, is there a respiratory system or heart or brain region of an artificial neural network?

Lex Fridman
Yeah, exactly. And I mean, if you think about science, right? A lot of scientific fields investigate things at many level of abstraction. In biology, you have molecular biology

studying proteins and molecules and so on, and they have cellular biology, and then you have histology studying tissues, and then you have anatomy, and then you have zoology, and then you have ecology. And so you have many, many levels of abstraction or physics, maybe you have a physics of individual particles, and then statistical physics gives you thermodynamics and things like this. And so you often have different levels of abstraction. And I think that right now we have mechanistic interpretability, if it succeeds, is sort of like a microbiology of neural networks, but we want something more like anatomy. And a question you might ask is, "Why can't you just go there directly?" And I think the answer is superstition, at least in significant part. It's that it's actually very hard to see this macroscopic structure without first sort of breaking down the microscopic structure in the right way and then studying how it connects together. But I'm hopeful that there is going to be something much larger than features and circuits and that we're going to be able to have a story that involves much bigger things. And then you can sort of study in detail the parts you care about.

Chris Olah

I suppose, in your biology, like a psychologist or a psychiatrist of a neural network.

Lex Fridman

And I think that the beautiful thing would be if we could go and rather than having disparate fields for those two things, if you could build a bridge between them, such that you could go and have all of your higher level distractions be grounded very firmly in this very solid, more rigorous, ideally foundation.

Chris Olah

What do you think is the difference between the human brain, the biological neural network and the artificial neural network?

Lex Fridman

Well, the neuroscientists have a much harder job than us. Sometimes I just count my blessings by how much easier my job is than the neuroscientists. So we can record from all the neurons. We can do that on arbitrary amounts of data. The neurons don't change while you're doing that, by the way. You can go and ablate neurons, you can edit the connections and so on, and then you can undo those changes. That's pretty great. You can intervene on any neuron and force it active and see what happens. You know which neurons are connected to everything. Neuroscientists want to get the connectome, we have the connectome and we have it for much bigger than C. elegans. And then not only do we have the connectome, we know which neurons excite or inhibit each other, right? It's not just that we know the binary mask, we know the weights. We can take gradients, we know computationally what each neuron does. I don't know. The list goes on and on. We just have so many advantages over neuroscientists. And then despite having all those advantages, it's really hard. And so one thing I do sometimes think is like, "Gosh, if it's this hard for us, it seems impossible under the constraints of neuroscience or near impossible." I don't know. Maybe part of me is I've got a few neuroscientists on my team, maybe I'm sort of like, "Ah, the neuroscientists. Maybe some of them would like to have an easier problem that's still very hard, and they could come and work on neural networks. And then after we figure out things in sort of the easy little pond of trying to understand neural networks, which is still very hard, then we could go back to biological neuroscience."

Chris Olah
I love what you've written about the goal of MechInterp research as two goals, safety and beauty. So can you talk about the beauty side of things?

Lex Fridman
Yeah. So there's this funny thing where I think some people are kind of disappointed by neural networks, I think, where they're like, "Ah, neural networks, it's just these simple rules. Then you just do a bunch of engineering to scale it up and it works really well. And where's the complex ideas? This isn't a very nice, beautiful scientific result." And I sometimes think when people say that, I picture them being like, "Evolution is so boring. It's just a bunch of simple rules. And you run evolution for a long time and you get biology. What a sucky way for biology to have turned out. Where's the complex rules?" But the beauty is that the simplicity generates complexity. Biology has these simple rules and it gives rise to all the life and ecosystems that we see around us. All the beauty of nature, that all just comes from evolution and from something very simple in evolution. And similarly, I think that neural networks build, create enormous complexity and beauty inside and structure inside themselves that people generally don't look at and don't try to understand because it's hard to understand. But I think that there is an incredibly rich structure to be discovered inside neural networks, a lot of very deep beauty if we're just willing to take the time to go and see it and understand it.

Chris Olah
Yeah, I love Mechinterp. The feeling like we are understanding or getting glimpses of understanding the magic that's going on inside is really wonderful.

Lex Fridman
It feels to me like one of the questions that's just calling out to be asked, and I'm sort of, I mean a lot of people are thinking about this, but I'm often surprised that not more are is how is it that we don't know how to create computer systems that can do these things? And yet we have these amazing systems that we don't know how to directly create computer programs that can do these things, but these neural networks can do all these amazing things. And it just feels like that is obviously the question that is calling out to be answered. If you have any degree of curiosity, it's like, "How is it that humanity now has these artifacts that can do these things that we don't know how to do?"

Chris Olah
Yeah. I love the image of the circus reaching towards the light of the objective function.

Lex Fridman
Yeah, it's this organic thing that we've grown and we have no idea what we've grown.

Chris Olah
Well, thank you for working on safety, and thank you for appreciating the beauty of the things you discover. And thank you for talking today, Chris, this was wonderful.

Lex Fridman
Thank you for taking the time to chat as well.

Chris Olah
Thanks for listening to this conversation with Chris Ola and before that, with Dario Amodei and Amanda Askell. To support this podcast, please check out our sponsors in the description. And now let me leave you with some words from Alan Watts. "The only way to make sense out of change is to plunge into it, move with it, and join the dance." Thank you for listening and hope to see you next time.