Stanford eCorner

'Helpful, Honest, Harmless' AI [Entire Talk]

23-02-2024

URL: https://ecorner.stanford.edu/videos/helpful-honest-harmless-ai-entire-talk/

Daniela Amodei is president and co-founder of Anthropic, an AI safety and research company. Amodei manages the senior leadership team, leveraging her people and management experience to further the company's goal of building reliable, interpretable, and steerable AI systems. In this conversation with Stanford adjunct lecturer Emily Ma, Amodei talks about how Anthropic's team of co-founders have built values and ethical guardrails into their AI tools from day one, and how this can inform other AI companies.

Transcript

(cheerful music) - Hello everyone. 00:00:17,430 Welcome back to the Entrepreneurial Thought Leaders Series, the Stanford Seminar for Aspiring Entrepreneurs. ETL is presented by STVP, the Stanford Engineering Entrepreneurship Center, and BASIS, the Business Association of Stanford Entrepreneurial Students. My name is Emily Ma and I head up special projects in real estate and workplace services at Google. Today I'm so excited to have Daniela with me from Anthropic. Daniela Amodei is the President and Co-Founder of Anthropic, an AI safety and research company. She manages the senior leadership team, leveraging her people and management experience to further the company's goals of building reliable, interpretable, steerable AI systems. Before co-founding Anthropic in 2020, Daniela was the VP of Safety and Policy at OpenAI and managed the people and research engineering teams prior to that. Even before that, she was part of the early team at Stripe where she managed recruiting and later led risk operations. She previously worked in international development and served as a congressional staffer.

Daniela and her brother, Anthropic CEO and Co-Founder Dario, were named to Times' 100 Most Influential People in Artificial Intelligence in 2023. She received her BA in English Literature, Politics, and Music at the University of California Santa Cruz. Everyone, please have a huge round of applause to Daniela on Valentine's Day. (audience applauding) - Thank you. 00:01:42,420 - Alright, well let's start with a big question. 00:01:46,260 So you are the first of this winter quarter who is a co-founder. And we've had, over the course of the last 25 years in this auditorium, a number of co-founders visit and share their stories. And you have a particularly interesting setup with your brother. So how does that work? - That's right, that's great. 00:02:05,796 Well, first of all, thank you so much for having me, everybody.

It's great to be here. And actually, to make it even more complicated than just being a co-founder with my brother, I actually have five other co-founders. So there were seven of us total who left OpenAI together. We all co-founded Anthropic. So if you're looking for a family feud that's definitely even more interesting. But Dario and I worked the most closely together of the co-founders as CEO and president. And I think I had a kind of great training wheels because I saw John and Patrick Collison at Stripe work together as siblings really kind of in the early days there. But I think what has worked especially well for us is just a ton of clarity on what are our zones of kind of ownership, competence and genius. And that's made it really easy for us to make decisions together as a unit. - So what would you say are your competencies 00:02:59,640 and what would you say are Dario's competencies? - So if you Google CEO and President of like, 00:03:04,560 a public company, it's almost a

perfect description of how Dario and I split our duties.

So Dario is the kind of technical, actually all my co-founders basically are technical except for me, and one other. But he really is thinking about what is the kind of five year arc of what's happening in generative AI? So much of the work that he and the team did at OpenAI was building GPT-2, GPT-3, but also a ton of sort of seminal work in technical safety research. And so much of his eye is thinking about, where is the industry as a whole going? What's happening in the kind of generative AI field? What's the research strategy and direction? I'm much more sort of practical kind of the business oriented side, but really work closely with the full senior leadership team and think a lot about how to actually grow and scale the company of Anthropic and then also what's happening on the product and business side. - I found it super interesting that, 00:04:01,093 while you're leading a technical firm and have been in technology for so many years, you actually have a philosophy and music and humanities background. So how does that work? - Yeah, it's so funny. 00:04:15,780 I mean, I think if I could have like looked forward into my future and seen what job I had, like I just wouldn't have figured out kind of how I wound up here. I think I sort of just tried to take the next best step in my career as I was kind of going along. But I do think something that's interesting is the humanities are a kind of great way to think about not just the technology itself, which you're building, of course requires incredible, you know, research and engineering skills. But so much of what happens in a business, and I saw this at Stripe and OpenAI and Anthropic and certainly in politics, is really about like, humans and how we relate to each other. And so much of how I think we scale these businesses, particularly in startups, right, where everything is new, you're kind of building the ship or the plane as it's taking off, right? To use the kind of overused metaphor.

So much of that is figuring out how to relate to, manage, coach, and lead people. And while that's not necessarily only the province of the humanities, it's sort of much easier to kind of talk about things in the people space when you sort of come from that background. - Yeah, for sure, for sure. 00:05:22,590 I was an engineering major here and I realized that the soft skills were so important to building product, building businesses. If I dial back and you could share one thing with a room full of engineering students, truly, what would you say to them to develop, maybe, that people side and their skills in leading people? - That's a great question. 00:05:47,160 I think, I mean, I heard that GSB has some like, touchy feely, I've heard about this from colleagues, yeah, something. Anyway, I have no idea if it's good or not. But like, maybe we're thinking about it. I think just in general, like just figuring out things about how you are and how you like to work. I feel like that sounds so obvious, but self-awareness is just one of the most useful skills in business, right? We talked a little before about the fact that, you know, I worked closely with Claire Hughes Johnson at Stripe.

And her book is amazing, just to give it a shout out, "Scaling People". But I think something I always really sort of admired about her and I think was a useful framework and learning for me is the more you understand about like, how you work and what bothers you or triggers you and what you are great at and what types of people you like to work with, you'll naturally seek those people out in an organization and kind of partner with them, which is a great sort of foundational building block for figuring out how to be effective in a company or with co-founders if you start something yourself. - So to get really specific, I am very, very, very impressed 00:06:54,510 that you manage technical people, right? And you've managed technical people. You've managed to get them to allow you to lead and to follow you. So how do you bridge the gap when you're not necessarily in the details of a technical project, or how an architecture is executed on? And I'm asking this question because I do the same thing myself, and so, but I'm an engineer. So how do you do that? - Sure. 00:07:22,140 I think it, first of all it doesn't work in every case, right? It has to sort of be the right setup, the right team. You have to have a technical person that is a strong technical visionary because of course I can't set that vision, right? I can't tell the researchers, hey, here's how to go, you know, train Claude 2 in the best ways. I just don't know that. But I do think there is a way that like, managing different groups of people, there's sort of are themes to it, right? Whether you're managing a giant sales organization or a group of researchers or engineers, you fundamentally are trying to get that group of people to accomplish something, right? You have to figure out what the problem is.

You have to figure out how many resources does it take to fix that problem or what is the goal and how will you know if you're successful? And so some of the process of kind of leading technical people doesn't necessarily feel that different than leading salespeople. The challenge is knowing what you don't know and sort of who to go to, right? When two kind of technical leaders or just strong technical individual contributors who were reporting to me when I was managing the language team or interpretability, the challenge was sort of figuring out, okay, who's right? If there's like, a sort of a technical disagreement. But my default there was, you guys are gonna have to talk about it and figure it out either way, right? So I probably shouldn't even be the arbiter. I should just sort of be supporting you in figuring that out. - Amazing, amazing. 00:08:51,480 So I wanna go back as well to self-awareness. Actually, I think there was a survey of GSB alums and 20 years out, 30 years out, the common sort of identification of a theme that they all said was actually self-awareness being the most important thing. So how do you yourself become more self-aware? - Yeah, that's great. 00:09:14,910 I think there's some sort of, these are, I would almost describe them as like, boring or practical ways to sort of support that, which businesses do so well. And I think startups in particular do this really well.

Which is we do performance reviews, which is like the least sexy sounding way to gain self-awareness I can possibly say. - I think it's very sexy. 00:09:36,720 - But really there's something about 00:09:38,730 when you work with people really closely, they actually know you really well, right? And I've had something like 12 or 15 years worth of performance reviews, which are a huge gift. But the shocking thing is like the content of those reviews, like the specifics and the details have certainly changed and sort of ebbed and flowed over the years, but the general themes are not that different, right? People are very

strongly who they are. And people that work closely with you will see those things about you and comment on them. So a way of cheating to get self-awareness is to just ask people that are close to you. My husband probably doesn't say things that differently from you know, my co-founders, of what is Daniela great at? And what are things that she's less good at? And what is she like on a good day and how annoying is she on a bad day, right? They'll probably all answer those questions with pretty similar themes. - Yeah, so what are you good at 00:10:33,870 and what are you bad at? What would they say? - I think the things the things in particular 00:10:43,170 that my reports say about me is I simultaneously really, I'm a supportive manager. I really want people to succeed. And I also hold a high bar.

Those are kind of the two things I hear the most. Also, ability to lead in lots of areas. That's another theme that comes up. In terms of things I'm less good at, probably unsurprisingly, I'm not the expert in anything, which is actually like-- - A feature, not a bug. 00:11:04,050 - Yeah, but it can be a drawback, right? 00:11:06,000 There's times where if something is, if there's contention between two parts of the organization, right? Product thinks we should do one thing, engineering thinks we should do something else, sales thinks we should do something else, I'm not necessarily deep enough in any of those like, disciplines to sort of say, hey, I think sales is missing this, or I think that the product team sort of has this right. So a lot of kind of general management versus specialized. And there are times in the business where you need an expert and I think really, the opportunity for me is always finding the right expert. - Ah, yes, very good. 00:11:37,350 Well, I know I've delayed, delayed to talk about Anthropic and I wanna dive in really deeply now. So tell us more about Anthropic.

Tell us about what the foundational philosophical ethic is of the company. You've innovated some really interesting things on how to make a responsive, explainable, steerable AI. Maybe we just start with an open book, like share with the audience what Anthropic is all about. - Sure. 00:12:01,680 So really Anthropic is aiming to build transformative general AI systems that are safe, reliable, steerable, ethical. And that's kind of a, it's a lofty goal, right? So much of, I think, why we were founded was this desire to build generative AI tools and systems and products that people could use while feeling really, really confident that what we were putting on the market was trustworthy, reliable, and safe. And that's taken a number of different forms, you know, over the past few years. But I think specifically we have this sort of belief that incorporating these technical safety streams into training of the model really from day one is the best way to ensure that when they actually are in a product form and getting into the hands of customers, that they're gonna be safe. - I love how your team has thought about using 00:12:59,660 the UN Human Rights Declaration as a starting point for how to really moderate content and moderate the algorithms development. Can you speak to that a little bit? - Definitely.

00:13:13,470 So this kind of idea that we came up with or our brilliant research team came up with is this idea of giving Claude, which is our generative AI tool, something we call a constitution, it's called constitutional AI. And to just sort of go back a little bit to kind of how we got there, the way that you sort of used to, by used to I mean like three years ago-- - Today still. 00:13:38,850 to make sure, hey, they're not sort of saying nasty things or they're not replicating bad inherent biases, was you would do this technique called reinforcement learning from human feedback, which we also kind of co-worked on when we were at OpenAI. And reinforcement learning from human feedback is just essentially a way of giving the model grades, right? It's like A plus, you said something nice, D minus, you said something mean. And that was reasonably effective in changing some behavior of the models. But what we found was that a lot of sort of subtler things about how models respond, react, things that they're sort of deeply believing are much harder to train out in those sort of individual cases. And so we had this idea of giving the model kind of a broader constitution in the way that you would in a society to basically say, what are ways of behaving and engaging that are good for humans, for people, and that don't perpetuate some of the problematic things that might be in training data. And so, so much of what we saw with constitutional AI was, A, we shouldn't necessarily be the arbiters of like, what is good or bad, right? We're a group of, at the time 100, 150 people, now 350 people based in San Francisco. Let's think about what broader documents that already exist in the world that have grappled with some of these challenges, and incorporate, I think we have something like 17 different founding documents in the constitution. - Amazing.

00:15:07,050 You know, I find it super interesting that you've chosen, sounds like a number of global documents. And so one of the questions that come up is society is very, very diverse. And so how do we adjust for specific cultures or ethnic orientations of what good and bad is? And how do you think about that as a team? Because it's complicated, right? You know, how Claude might respond to, let's just take myself, an Asian American born in Canada, might be very different in conversation with someone who was born in Taiwan who has a very different maybe view of sort of humanity and morality. So how do you think about those challenges? - Yes, this is a great question. 00:15:49,800 And it's something our teams think about a lot, right? For sort of the work that we do. So I think constitutional AI is sort of a great framework for thinking about what are the inputs that you would want to put into a model to kind of guide its ethics. But those don't necessarily, those inputs can change, right? So depending on the culture, the country, the company even that's using our models, there's probably some degree of latitude over time that we can build in to say, hey, do you wanna change sort of some of the founding documents of the model that you use? That being said, I do think there are some guardrails that we feel strongly should be in place, you know, regardless, right? Claude can't help people create weapons, right? It's like trained to ensure that you don't do something harmful to people or animals. And I think sort of regardless of where it's deployed, those are kind of company universal things that we feel, you know, believe strongly in. Another thing I'll point to is we recently did a set of research that was kind of building on this constitutional AI idea called collective constitutional AI. And what we did there was instead of using just a set of founding documents, we actually polled a very large, kind of demographically diverse group of people from sort of around the country and in other parts of the world to see like, how did they react to some of the things that were in our constitution and what did they kind of collectively come to on some of these sort of ethical questions? And how far was

the distance between sort of what our constitution said and what the collective constitution said? - Hm, wow.

00:17:21,453 So more broadly then, with the ethic that you have at Anthropic, how do you see your role in the broader, societal efforts around governing AI in a responsible way? I mean, I was thinking about how we kind of have one chance to get it right and we can't let people down. So how do you see your role, Anthropic's role in helping to guide what AI could do for humanity? There's so much opportunity, yet there is risk. - Absolutely. 00:17:56,330 So I think what's interesting about this is, you know, Anthropic is a company, right? We're a public benefit corporation and we have this very lofty social mission. And I don't think anybody, including us, thinks it's right for sort of us or any particular company alone to be the arbiter of what happens with this technology. And so there's a few different ways that we try and collaborate or work with other groups to say like, where does this decision live? Who is kind of the group or the set of people sort of driving these outcomes? And one way that we do that is we've had a policy team basically since day one. The only other non-technical co-founder of ours is our policy director. And he has just done an incredible job working with policymakers and government officials to really think about, you know, what are the ways that we want this technology to be regulated? Or what are the rules and guardrails sort of beyond the corporate level where we might want this technology to be looked at, right? What are the sort of insights about it that we need to share to the government and policy makers around the world to help them understand what's happening? We also work with a number of civil society and nonprofit groups that work on many of the issues that you've talked about. And that's because those groups often have additional expertise that we don't necessarily have within our walls, right? There are groups that think specifically about particular elements of how this technology will be used or abused, where we can lean on their expertise. - Ah, very good, very good.

00:19:36,660 Building on that, on the flip side, as a user of these tools, it's not always clear. So I spend a lot of time helping nonprofit organizations understand the potential of AI, but also the risks, right? And it's not always apparent that there's some challenges with these tools. It's not always accurate. Large language models can hallucinate. Depending on the data you provided, there could be bias. There could be essentially people unfortunately putting private information into public models. How do you see Anthropic's role or your role or our role as citizens in understanding how we need to show up to be a good partner to these tools? - Yeah, I think I have sort of a like, short term answer, 00:20:22,890 which is what do we do with kind of what is available to us today? And then kind of a like more speculative kind of long term answer. So I think on this sort of short term front, I think there is a lot of just really interesting work being done around some of these fundamental questions, right? What does it mean to sort of use these models and understand what's happening inside them? And we try to publish our research about all of this, right? We don't have perfect information because these models are, even to the people that are training them, still a little bit of a mystery, right? We know, hey, you put in data, you put in compute, you do some fancy magical algorithms and like magic, right, you have these really powerful tools. But all of the sort of details underneath are a little bit opaque still. And so I think to the degree that we are kind of able to, whether it's with you know, customers or lawmakers or individuals, sort of explaining what we've done is kind of really a big part of the ethos of Anthropic.

In the longer term, I sort of particularly want to click into a research team at Anthropic in the area of mechanistic interpretability. Which is an area that, again, one of my other co-founders sort of pioneered, first at Google actually, and then at OpenAI and now at Anthropic. And really the best way to think of mechanistic interpretability is almost as the equivalent of like, neuroscience and what neuroscientists do to the human brain, what mechanistic interpretability experts do to neural networks. And so really thinking about when models, when these sort of neural networks are producing outputs, we don't know what's happening, right? Just like when humans are sort of thinking, oh, I think this thing about this person, why do I think that? If we could actually go in and say, what are the literal neurons that are firing that are causing the model to think this or do this? And are there combinations that are maybe problematic that are firing together, right? And so even if you can sort of train it out of the model at the end, it would be much better if we saw, are there kind of problematic things, or positive things happening in the model that we would want to adjust. - Oh, that's fascinating. 00:22:29,220 It's almost like doing brain surgery on a neural network. - Exactly, yeah. 00:22:38,580 - So Claude 2 is a wonderful tool. 00:22:39,660 You should all check it out. You can try it at claude.ai.

And really so much of how we have been using and thinking about sort of generative AI and sort of Claude is as a tool that has just a wide variety of different applications. So we offer a first party API, and so this is for developers to be able to build on, also some larger customers also work with us through our first party. Claude AI, you can play around with on your phone, you can try it on your computer, very general assistant. But in general, the thing we've kind of heard it's best at, or people most prefer using it for, is long content. You know, writing long things, right? You can upload up to 200,000 tokens of context, so that's about two books worth of information. Claude is great at summarizing or pulling things out. I'm definitely not telling you if you have an English class to cheat on your homework. That's not what I'm saying. - Yeah, don't cheat. 00:23:41,670 if you want Claude to sound like you and producing content or writing works of fiction and things like that.

Or nonfiction, but don't use it for school obviously. (both laughing) - You heard all that? 00:23:49,859 Just triple checking. So what are some of the other things that you've seen people use Claude for that have surprised you maybe? - Sure. 00:23:57,210 I think maybe one of the most surprising kind of macro trends that we've seen is actually very, like large enterprise businesses have been some of our earliest adopters. And that's generally not the kind of sort of market adoption trajectory that most businesses see. Part of why I think that's happened is because Claude and Anthropic kind of have this reliable, sort of trustworthy, scalable set of values and kind of our approach to training the models. More kind of traditional industries like, you know, insurance, healthcare, legal services, right? These are the types of industries that really value

reliability, trustworthiness with their end customers. And where things like, you know, lower hallucination rates are actually a huge deal for them. Claude has the lowest hallucination rate on market. And so a lot of what we've seen is traditionally kind of companies that might not be the first to sort of adopt a new technology are actually some of our kind of biggest adopters.

- Huh. 00:25:06,180 And could you maybe share an example of how a healthcare company might be using Claude right now? - Definitely. 00:25:10,583 So I'll go through a few examples. So on the kind of financial services side, groups like Bridgewater are using us for financial analysis of their tools. We have a mortgage lender that basically has a huge amount of data and is using it to help people fill out home mortgage applications and shorten the time that it takes for them to be approved. We also work with another financial services company that's using it to sort of help people figure out how to do their taxes better. In the healthcare space, what's really interesting is Claude can be a great partner in concert with a medical practitioner, so a doctor or a nurse, and really going through and for example summarizing key findings from a health consult, right? You can't use Claude alone today, but it often saves doctors and nurses a huge amount of just like, administrative time, right? So much of what we hear from doctors and medical practitioners is, I would love to spend more time with patients and less time doing paperwork. And Claude is a great partner for helping them to not do as much paperwork. - Yes. 00:26:11,670 So on that note, what do you think about agents and AI agents built on top of LLMs? - It's a great question.

00:26:20,670 I think this is a place where there's sort of two things that are true. I think long term, I imagine this being like an incredibly powerful technology that will save many people time and labor and headache and administrative burden. I also think it's interesting that today I don't think the models are quite there. - I agree. 00:27:10,200 And I think there is a way that sort of, it feels when you sort of look at it and you're oh my god, it's like a human, it's like talking to a person, it can do anything. It's not there yet. There's still many things that humans cannot use generative AI to even help them with. And most applications of generative AI today I actually think are best when done with a human in the loop. Not all of them, but the majority of them. - Well said, well said.

00:27:35,463 So what are your dreams for Claude 3? - It's a great question. 00:27:40,170 I think for Claude 3, number one, on some of the just core kind of safety features, right? We have this kind of helpful, honest, harmless framework. We're hoping to just make improvements on all of those. A huge one is, the number one request we hear from model quality from our customers is how do you get the hallucination rate from, you know, like X percent to 0%? Like whatever the number is, nobody wants a model that's going to make up information. And while we think we've seen fairly impressive gains there, it's really hard to get to zero. And so I think we're always kind of chasing that number. Also, some interesting questions around like getting the models to just hedge a little bit more if they think they don't know the answer. I think on the harmlessness front, this is again a place where Claude is sort of industry leading on this, but there's always more work to do and sort of understand. And then I think another set of kind of features that we're sort of interested in for Claude 3 is really how do you make the model more generally capable, right? Just sort of more intelligent. But really how do you get it to sort of specialize in particular use cases or industries? - I really appreciate 00:28:45,990 what you said about the two ends of the spectrum.

On one end you wanna get the hallucination rate down to zero, which is almost impossible. But on the other end, you wanna make sure that Claude is still useful. And I have found that with some other models we've kind of swung to the conservative end, right? Like I remember in the beginning days with Gemini it would just make stuff up, but it would give you an answer to anything, right? But now it will deliberately not answer who is Emily Ma, right? Because it's ambiguous who I am, truly, when you scour the internet. And so where do we sit on that spectrum? Like, we wanna be responsible with large language models. How do we approach that in the most honest and transparent way? - So I'll start with kind of a funny story here 00:29:38,040 to sort of illustrate what you're talking about, which is in sort of early days of training Claude, we were really experimenting with sort of trading off some of these H's, right? This like helpful, honest, harmless. You can have a perfectly harmless model if you want it, it would just not be very helpful, right? Like we would just sort of ask Claude like, "Who is the first president of the United States?" It would be like, "I cannot answer that question and I'm also very concerned about your wellbeing. And here is a link to like, you know, a harm prevention website." And you're just like, "Claude, I'm fine, I promise I'm fine." So there there is sort of this chart of sort of intersection, right? Where you can say, okay, do we want Claude to risk a little bit more helpfulness or a little bit more honesty, right, or a little more harmfulness, or however you might describe it? And of course what we're always trying to do is sort of raise the watermark on all three. But at the end of the day, depending on the application, you might also want to sort of fine tune the model or train a different model for certain use cases, right? You can imagine that for an educational tool for like, you know, five to eight year olds, you might want a very, very harmless version of Claude even if it's like, a little bit less helpful. Whereas if you're using Claude to do, for example, trust and safety detection work, which is an application that many of our customers use Claude for, so that humans don't have to look through harmful content, you actually want Claude to be able to read and identify and understand very harmful or upsetting things and sort of filter them out. - That's right, that's right.

00:31:11,100 So I wanna take the last couple of minutes in this sort of more formal section to talk a little bit about how Anthropic came to be. So going back to the origin story. And one of the things that we grapple with as a class is what are the principles by which we live by and how do we transition well from one to another, right? And it was very clear that you and your co-founders had a vision for what could be and that the sort of circumstances that you had at OpenAI weren't going to allow you to manifest that. How do you gracefully say no, so to speak, and close out one chapter well, and then start another chapter? That's a skill I feel very strongly about as entrepreneurs that we need to get good at. Whether it's when we see a product that isn't actually, you know, being adopted as much and quickly as we want and we decide to pivot, or it's constantly

we're iterating on our lives and having to not hold onto things for too long and be able to move forward. So how did you approach that situation, that circumstance with Dario and with your other co-founders? - I think what was probably the most interesting 00:32:23,130 about that was when you sort of have something that you're running towards, it makes it much easier to feel grounded in your values. We felt very strongly, my six co-founders and I, we were mostly the kind of technical safety and policy leadership of the company. And something we felt very strongly about was this kind of vision of building transformative AI systems in a way that was reliable and safe and transparent. And that was like a theme that already was very live for us because that's so much of what we worked on. And so I think in a lot of ways, it just felt like a natural next progression, right? We said, "We have this incredible vision, we really want to be able to go and do it." And I think that sort of, it's almost like in any, you know, relationship or situation when you see something that you're like, this is really what I'm meant to go do next, there can be quite a lot of internal clarity that can kind of drive you forward.

- Okay, fantastic, fantastic. 00:33:19,470 Well, I'm gonna ask you one more question and I'm gonna hand it to the classroom to ask many, many questions 'cause I think they're all just bubbling, ready to go. So if you were sitting here however many years ago when you were 20 years old and you could tell yourself something, what would you tell that Daniela sitting in the front row, the 20-year-old version of yourself? - Oh my gosh. 00:33:47,730 I think I would say, number one, follow your passion. And normally when people hear that they're like, oh, I like soccer, I should go be a soccer player. Like whatever is the thing that is like, most exciting and drawing for you now, right? I didn't have sort of a traditional career trajectory of saying, I'm gonna study engineering and then I'm gonna go work at a tech company. That was just not how I started out. And I truly think that the time I spent working in international development and global health, and I worked on a campaign and I worked in politics, I learned so much about myself and what I like to do and also what I'm good at. And also very importantly, what I don't like, right? I had an incredible opportunity to work on Capitol Hill and I was like, this is extremely not for me, right? It just was slow and bureaucratic and I missed the campaign days of getting to build something together with a small group of, you know, motivated people. And that was such a like, light bulb moment for me when I essentially was like, the technology industry and the startup world in particularly is that, right? It's a campaign but sort of in a long term, slightly more sustainable way of doing it.

So I think, I mean it sounds so cheesy, but just like, follow what is exciting for you because when you find something you love doing, you can work at it very hard for a long time. And when you hate doing something, 10 hours a week of it is miserable, let alone 40. - So follow your passion. 00:35:15,030 - And know your passions might change. 00:35:17,280 That's the other thing, like when your passions change, don't hold onto old ideas of what you liked to do. Be like, I thought I liked this, right? I saw myself working on Capitol Hill and I was like, oh my gosh, this is who I'm meant to be. And then I got there and I was like, wrong turn. - Oh, fascinating. 00:35:31,230 And also you weren't stuck on campaign specifically. You were able to sort of zoom out and say, these are the characteristics I loved about working on a campaign.

Where else could I find it? What other industries, what other areas in the world could I find that similar setup? - That's right. 00:35:59,610 to your investors while still having this strong mission of making safe AI. - That's a great and very timely question. 00:36:04,740 And I know this will shock you, but we're actually asked this question quite a lot now. So Anthropic, I think, sort of from day one has been very curious about this question. And we incorporated as a public benefit corporation exactly for this reason. And public benefit corporations are very, they're like a very interesting form of governance. I highly recommend, like go Google them. Or ask Claude. Ask Claude.

(Emily laughing) Public benefit corporations have been around for quite a while. They're a corporation, right? They're a C corp. But they have this additional component of a public benefit mission. And so like TOMS Shoes, Patagonia, these are public benefit corporations, highly profitable. And a lot of sort of what, the reason we sort of decided to go that way, we thought about a lot of different structures when we were starting Anthropic. We were like, should we be a nonprofit? Should we be an LLC? Should we be a C corp? And we landed on PBC because we felt that it was sort of best positioned to provide us this kind of flexibility of having investors, right? Like issuing equity, we're a normal company, we have products, but there's this social mission that sort of, we are somewhat legally protected from shareholder lawsuits. If we decide, for example, we don't know that Claude 7 is as safe as we want it to be, we're not gonna release it yet. Whereas in a traditional C corporation, you have a much bigger risk of that being a problem if your shareholders don't like what you've done. It's not perfect, but that's sort of a nice protection. Additionally, we have this group of people called the Long-Term Benefit Trust who are financially disinterested, so they're not investors, and they essentially elect a percentage of board seats that sit on Anthropic's traditional board of directors.

And the way that we selected them was for their experience and interest and influence in the public benefit sector. And the last one I'll say is we recently published something called a Responsible Scaling Plan. And this policy, essentially it's public, you can read it on our website, it says what will we do if we are concerned about sort of particular risks from either a technical, safety, or security side of the models that we're developing? How will we measure these and what will we do to sort of react if we're concerned that what we're going to put out into the world is gonna cause harm? Student Thank you so much for being here. 00:38:22,923 My question is, how did your time at Stripe, which is a massive FinTech company led by the Collison Brothers, sort of prepare you to take on leadership now yourself? And how did you kind of, I guess twice in your career, leaving Stripe and then also OpenAI and maybe other moments, have the confidence to like, leave things that were going great behind, or maybe they weren't going great, but like, to then do something that felt right to you? - So I think on the kind of question of like, 00:38:51,570 how did scaling at these other companies sort of help prepare for Anthropic? Something

that is interesting is all three of those companies are very different. But hyperscale, there's some themes about it that look really similar. And I think in particular, helping to scale Stripe from 40 to 1500 people and then I joined OpenAI around the same size, it's like, this is the third time I'm watching the movie. And so there are things of course that are completely unique to Anthropic, but there's like, oh, we've gone through this amount of scale, we have to add this set of processes because 300 people communicate differently than 30 people. That like, many of the folks I work with who have not done hyperscale before, like, what is happening? I don't know how to find information anymore. I'm like, totally normal, right? And I actually write a document at the beginning of every year saying, here's what we should expect from just a pure scaling perspective over the next year at Anthropic, because I've seen it before, right? And it's not perfect, but I think there's kind of similar themes. The other thing I learned was really, just I got to see such a wide view of all of those companies because I worked in so many different pockets of them that now at Anthropic I feel much more confident saying, you know what, I've never managed a sales team before, but I can probably figure it out.

And I was very grateful at both Stripe and OpenAI to be given just opportunities across different parts of the company. The second part of your question I think was, repeat the second part of your question for me. Yeah, how do you know when to leave? Yeah, I love this. So it's an art and everyone will do it differently, I think is my real answer. For me, I loved Stripe. Like, Stripe is a great company and it was a hard decision to leave. I think again, to sort of go back to this like, running towards something versus away from something, the earlier part of my career had really been about sort of wanting to have positive impacts in the world. And Stripe is a great company and payments is cool, but it didn't sort of feel like fully meeting this other aim or goal that I had around wanting to like, positively impact the world. And so when I saw the opportunity at OpenAI and I knew people there, I thought, wow, this feels like a really exciting, transformative technology, I kind of wanna go there. And then with Anthropic it was again very similar that I had this vision around safety and trustworthiness and felt very confident that I wanted to go do that next.

Student Hi Daniela, thank you for this talk. 00:41:30,210 So today the hot topic is AI. A couple of years ago it was NFT, crypto, Metaverse. What's the next hot topic in your opinion? And are you working on it now or not? Thank you. - If only I knew the future. 00:41:45,240 I would go found that company if I knew what it was. You know, I think maybe an interesting version of your question is like, is the sort of generative AI hype going to continue, right? Like, are we in a over exposure period, right? Like I see some heads nodding. We're like, you could only read about a technology for so long without sort of being like, okay, I get it. Like yay, it's changing the world. I think what's interesting is, you know, we've had a lot of splash and aha moments in artificial intelligence.

And I think there's a really interesting phase that we're entering now that I see particularly on the business side, which is companies, at least in 2023, were experimenting, right? They're like, how can I use this technology to remove obstacles or pain in my organization or make us more effective or improve search or understand my customers better or structure data differently? And it was just very exploratory. And I think what we're seeing in 2024 is just more kind of boring business scaling, right? We're like, okay, you've had access to this technology for six months or a year. And people are like, CFOs are now getting involved, being like, how much are we spending on this? And like, what's the actual bottom line return on it? And I think what we'll see is a little bit of a departure from just the like, it can do anything language, to sort of more detailed analyses of like, what exactly can the technology do today? And then every time there's a new model release, there will be more of a discussion around what can it do now that the last model couldn't do? Maybe out of that comes some other incredible technology that some other company will build. I don't know. But I think if I were to sort of make the case for what I expect to happen in AI in 2024, it would look like that. - Can I do a follow up question? 00:43:34,290 - Please. 00:43:53,160 to run all of these models. So what are your thoughts on that? - There's a really sort of interesting conversation 00:44:03,570 that I have a few times a week with a customer. So number one, there's so much excitement about AI, right? Which means Fortune 500 companies are like, knocking down our door, right? They're like, oh my gosh, gen AI. Like let us in, help us.

And I think in many cases there is a good use case. And we'll say, hey, it sounds like what you're looking for is help understanding complex data, right? And pulling out key pieces of information that would take regular search much longer and lower accuracy to do. Or you really wanna understand, analyze sentiment, right? That's much harder to do. But there are some customers who come to us and the honest conversation we have with them is, I don't think we can do anything for you right now. And we don't wanna sell you a bill of goods that we can to close a deal with you and get your logo. And I think there's a fundamental kind of honesty and integrity about that that feels very important to us. And so much of what we do there is, A, we try to leave a good taste in their mouth to feel like, we didn't lie to you, right? We didn't say, oh, AI? It can solve your staffing problems. Like, you can use it as a manager, it can't do those things yet. But I'm like, come back to us six months from now, right? Or what are the biggest problems and challenges in your business? Like, we'll write those down and we'll come back to you when we think that Claude can provide the solution for you. But I think some just being clear-eyed about what the technology can't do and being honest with businesses about that feels important.

- It's a great theme. 00:45:30,330 You know, helpful, harmless and honest not only applies to Claude, but how you run the company and how you approach your partners. - Employees. 00:45:36,298 Yeah, we're like are they helpful? (Emily laughing) Just kidding. - Are they harmless? 00:45:41,580 Couple more questions. Oh goodness. Student So this might be a sticky subject 00:45:46,770 but I'd love your answer. So AI is likely gonna be a very decisive technology in future military conflicts. And a lot of the requirements the DOD and various US government agencies have with regards to AI are a lot aligned with what you've talked about, like trustworthiness, transparency, et cetera. So is Anthropic open to working with the US

government and DOD? And if so, can you talk about how you would rationalize kind of this concept of doing no harm in the context of war? - Very important question, especially for a company 00:46:21,210 that's working in this sort of field that we're working in.

The way that we do this now is we divide whether or not a business can use Claude not by the business but by the use case. And so our acceptable use policy says, hey, if you wanna use Claude for like, processing backend employee records or something at an institution, if it's not sanctioned, if it's not a business that we can't work with for some reason, that's allowable. The place that we put restrictions are on what the use cases are. And so our current AUP says you cannot use Claude for things like military applications, for weapons, right? There was this interesting article you probably saw of like, simulation of like, you know, basically using these models to do war things. And Claude was the least escalatory. So Claude was like, let's not launch missiles, let's sort of downplay this. Claude is very nice. So that's obviously our hope for what we would be building into the models, but the way that we think about that today is through our acceptable use policy. Student Thank you so much for being here. 00:47:25,110 My question is on agents.

As you clearly said, the models are not quite there yet. Do you see in future Claude's sort of becoming this all encompassing agent that can book your flights and do everything? In terms of direction of Anthropic, are you sort of trying to go there? - So I think on the agents front, 00:47:45,360 there's almost like a question above it that I kind of wanna go to, which is are we building kind of tools that are sort of intending to do like particular tasks, or to kind of be like generalized sort of assistants? Are we trying to get them to do everything? And I think my real answer is like, I don't think the industry knows yet. And I think today it seems to me that we haven't quite crossed the chasm between these models being great at doing sort of a set of individual tasks or maybe a combination of tasks to it understanding the sort of full scope of like, general purpose project management or something. I think that's just today sort of a bridge too far for what they can do. That being said, I could imagine that like, Claude 6 or Claude 7 could do that, right? And a component of that is sort of a product one, which is how does the model kind of integrate with your Google Calendar and your flight booking and Chrome or whatever the sort of applications are? But there's also just like, a model capability component of it that I actually think is the tougher piece of it. My instinct is we're probably still a few generations from that being possible in sort of a general way, but there might be sort of narrower or simpler sets of tasks, right? Like making a reservation somewhere might be easier than booking a whole vacation. And so I think we'll probably see something like a, just like a curve or like a gradient of a tool like Claude being able to do that over time. But it's a great question. - Let's do one more question. 00:49:26,013 So lucky, lucky final question.

Student I know you mentioned a lot of 00:49:34,470 exciting use cases in companies and enterprises starting to adopt closed source APIs like Claude, OpenAI, Cohere, the rest of them. I mean, what do you think are the biggest struggles today for enterprise to continue to create emerging workflows and use cases with these technologies? And if you had to start an AI company facilitating this process for the enterprise specifically, what would it be? - So I think there's a few. 00:50:01,950 I think the first is depending on the kind of amount of technical capacity within that company, it can vary really widely. So most big enterprise companies have like, a technical team, but some are more technical than others. And so I think there is still a little bit of a dance that happens at the sort of like, integration and just getting the sort of model started working in the guts. Just like a very technical sort of side of it. That is still like, a process that is not nailed down, I think anywhere. I think like for any of the sort of LLM providers or generative AI providers on any cloud. I do think that's getting better just as we sort of have more data. But I actually, that's like the number one most difficult part is just like, plugging it into the right places to be able to use it the way you want, prompt it the way you want, and sort of teach your workforce to do it.

The second is, it's a little bit particular, but the sort of like, what are the data sources that you're giving to the LLM, the sort of structure of those can vary really widely. And so there's also a kind of component that's almost just like figuring out what it is you're trying to plug into the model. Not just technically, but sort of organizationally. You're like, we have to pull this thing from this system and there's this other set of data from this system that also is almost like just another hoop. And I think that it's interesting to imagine, maybe there's like a bridge technology there or a bridge company there that's like, we go and do all of that for you on the backend and sort of prepare you to just be plugged into one of the foundation model factories. That might also come from a company like us, but I think probably that's like, not the place that we're gonna have the most innovation. And I could definitely see an interesting business idea there. - I find it so interesting to end on this front. 00:51:39,030 For me, with gen AI, it actually still starts with the people, right? It starts with the input. The technology is only as good as the people and the inputs.

And so as we think about adopting gen AI more broadly, it's a question of, what is the business need? What do the people need? And how do we approach it from that front? So with that, I know all of you are rushing off to your next class, so let's give Daniela a huge round of applause. Thank you so much, so much. (all applauding) Quick note, next week back in this classroom we have Shiza Shahad, who is the CEO and Co-Founder of Our Place. She was also the co-founder and formerly founding CEO of Malala Fund. So for those of you who know Malala, she helped found the fund to fund education around the world. And if you're interested in these events, you can find it on Stanford eCorner on the YouTube channel, including this one next week, and lots of other videos and podcasts and whatnot. So thank you again. Huge round of applause. (all applauding) (mellow music)..