Stanford eCorner

Better AI Through Better Data [Entire Talk]

Alexandr Wang, *Scale AI*

26-01-2022

URL: https://ecorner.stanford.edu/videos/better-ai-through-better-data-entire-talk/

Alexandr Wang is the founder and CEO of Scale AI. He founded Scale while at MIT at the age of 19, after recognizing that he could accelerate companies' abilities to deploy AI by combining a machine learning-powered data labeling system with human insight to ensure that models are trained on high-quality, trusted datasets. In this conversation with Stanford adjunct lecturer Emily Ma, Wang discusses the essential role of high-quality data in building powerful, useful, and unbiased AI and machine learning systems.

Transcript

(gentle music) Female AI Voice Who you are defines how you build. 00:00:06,993 - Hello everyone. 00:00:07,930 My name is Emily Ma, and I'd love to welcome you to the Entrepreneurial Thought Leader series today, presented by the Stanford Technology Ventures Program, which is the entrepreneurship center in Stanford School of Engineering, and then BASES, the Business Association of Stanford Entrepreneurial Students. I am so excited to have Alex Wang at ETL today. It's also his birthday, secretly, and so he chose to spend his birthday with you, which is the coolest thing ever. Alex is the founder and CEO of Scale AI. He founded Scale AI at the age of 19 while he was studying at MIT. His central insight was, that machine learning powered data labeling could really help human beings make better use of AI over time. Under Alex's leadership, the Scale has grown to a $7 billion valuation and sourced hundreds of customers across so many industries from finance to e-commerce to U.S. government agencies.

Welcome Alex, let's start off with what Scale AI does. I mean, you've built this incredible company in three years. Tell us more about the breadth of work that you do. - First of all, thank you so much for having me Emily, 00:01:15,290 really excited to chat today, and I really enjoy chatting with you, so I'm excited. So yeah, Scale. So our mission is, to accelerate the development of AI applications. We believe that AI and machine learning is, if not the most important technology of today, one of the most important technologies, that's going to enable huge amount of goodness for the world, and just enable the world to operate significantly more efficiently, and effectively enable new customer experiences. And our vision in accomplishing that, is building the most data centric infrastructure platform for AI and machine learning. And the real insight that we have, or the thing that powers everything that we do, is this thought that data is the new code, and that the thing that will dictate the performance of these machine learning systems and these AI systems of the future is actually the datasets and the data that they're trained on, much more so than the code that is written to power them. And so if we were to boil it down to a two sentence, it's that better data results in better AI.

And we've taken that sort of core idea and used that to build out an infrastructure platform, to power a large number of, a large swath of this sort of like AI ecosystem or the sort of like AI use cases out there. So we originally started with data labeling, or data annotation. It was the problem of converting sort of raw data feeds to useful label tag data that we can actually use to train large-scale machine learning systems. We actually started in autonomous vehicles. And then since then,

we've scaled across a variety of different industries, like you mentioned, from e-commerce to financial services, to the government, to working with large tech platforms, and sort of everything in between. And what we do with these customers is we help them, not only with data annotation, data labeling, we help them with data management. We help them build out actually algorithm. And so with some customers, we provide algorithms directly to them for stuff like document automation, or e-commerce AI, or in the government use cases. And so we've been able to expand pretty rapidly into a huge number of product areas, into a huge number of verticals. But again, essentially at the core, it's all been powered by this concept that better data results in better AI, and that the most valuable thing that we can do to ensure that we have great AI systems of the future is to build incredible systems, to build great data sets.

So that's what we do. - God, that's amazing. 00:03:44,500 And the fact that, you came across the seedling of that insight when you were 19, you weren't even 20. And I'm actually curious if we just go back now, knowing what Scale AI has grown to, when you were 19, what gave you the confidence in this concept that data is a new code, that was the basis, that was the soil, for the growth of beautiful AI algorithms in the future? And what gave you the courage to jump in and do this? - Yeah, so, it's sort of a few things. 00:04:22,270 So when I was actually, I was at MIT, I was studying AI and machine learning. And this was the year when Google released TensorFlow, and DeepMind released AlphaGo. So sort of this like big seminal moment for AI and machine learning. And it really felt, it was this, I actually remember there was like this reporter, who was walking around MIT campus and was like interviewing MIT students to see what they thought about AlphaGo. And so it was like, it was this very clear moment where I felt like, hey, AI is actually is going to happen. It's gonna be big.

And then I remember, both in some projects, as well as like in some school projects as well as some side projects, I, in one of these side projects I remember very viscerally, I was like, I wanna build a camera inside my semi fridge that would tell me when my roommates were stealing my food. And I remember very viscerally. It's like, hey, there's all these great neural networks. They're really, really cool. But at the end of the day, the algorithm is only as good as the data that it's trained on. And so it was like, hey, this is going to be a critical, almost pillar of whatever AI looks like in the future. And I looked around and I realized like, hey, this is a big problem that there aren't actually that many people trying to solve, or there aren't that many people focused on solving this problem. And ultimately, the thing that gave me conviction was frankly, I had seen sort of the success stories in sort of the years prior, of platforms like AWS, which enabled everybody to build these large-scale internet systems, or sort of like websites, and large-scale internet platforms. I've seen the success with platforms like Stripe, for it to enable payments and enable sort of like, you to build businesses on the internet. And so ultimately, the realization was kind of, hey, you know what AWS has done for the cloud or what Stripe has done for payments, there's an opportunity for a company to do that for AI and unlock this huge amount of potential for the technology by solving one of these critical pillars.

And the pattern recognition was that, hey, data centric AI was going to be just as important as some of these other sort of like foundational pillars. And so that's really, what kind of got me excited at the time. And, honestly speaking, I didn't have all the answers. I didn't know necessarily, that this idea was definitely going to be, as important as I think we believe it is now. Or I didn't necessarily know that it was going to be, as exciting as we think it is now. But the sort of fundamentals were there for it to be like, hey, this is certainly worth exploration. - I have a very important question for you. 00:07:09,930 Did your roommate stop stealing your food? - Well, the punchline is actually, I couldn't build it, 00:07:15,960 because just building that would have required so much data about different foods and like different configurations of what it looked like to pull food away from the fridge. (Emily laughs) It was just like, literally the amount of data would have been astronomical that like, it took me all of two hours to just give up on that idea and realize, there's no way. - And building on this question though, 00:07:35,190 there's more, seriously, under that, so, I hope they stopped stealing your food, regardless.

You actually had, some other entrepreneurial sort of experiences. You had some other things that you were building with classmates, and I'm curious if that sort of helped you feel out what the entrepreneurial journey was like. - Yeah, I think that there's like, 00:07:58,140 I often tell, I tell everyone this, which is that I think this sort of like, this creation process of having an idea in your mind, and then working with people to like, realize that idea into the world, it's an incredibly empowering process. It doesn't always work, obviously. It's not always the case that like, you build something and it's amazing. But I think it's like, it gives you so much confidence to like go from, literally, an idea that you're dreaming up and you're sort of like working on in iso- You're literally dreaming it up to go through this sort of like process of making that a reality, and identifying what are all the components we need to build. How are we gonna architect it? How do we build that? And actually build it into a reality. 'Cause it gives you so much, it gives you so much power as an individual. It's like, wow, I can, I have this ability to like, will things, from my mind into my, into reality. And I think that, and I think that's like one of the most valuable things, that you can gift to anyone, right? And we actually, we have a value, like it's at Scale, we call it ambition shapes reality.

And maybe one day I'll sort of publish, like, how we think about this internally. But the core idea is that like, there's this incredible quality where if you, doing something that's ambitious versus not ambitious, they might take literally the same amount of work. And so, if you are able to like dream up big ideas and dream up big solutions to problems, and then you work to build them, you can have just like this incredible sort of like, a feedback loop with the world in building great things, and continuing to sort of like expand your horizons. - Wow, I love it so much. 00:09:44,530 I think there's two things that I took away from what you just said. I think a lot of times in, especially as students, right? It's like, hey, you know what? I don't have the skills to build that yet. But like, sometimes it's really powerful, just get going, and to realize that you can manifest, right? You can code, you can build an app, even if it's not perfect, right? You can manifest from an idea in your head into something,

right? And just that prototype is like very empowering. It shows that we have that agency to manifest, manifest and manifest. And then the second thing that you mentioned, actually made me sort of think about like my own journey in food. It's like, wow, you know what? If I aim for like 100% better, I'm only gonna get maybe 80% there, lucky, maybe I get 100%.

But aim for like 1000% better. I might get 500%, right? So, if we aim further with our ambition. We actually get further. If we set our finish line too close, it's too easy. We actually shortchange ourselves in many ways. So I hope you write that book. I really do hope you write that book 'cause it's a powerful message. - It's so true. 00:10:43,980 There's all these examples. I collect examples of this in real life.

And one of the best ones is that like, the four minute mile. For the longest time, humans thought that the four minute mile would have been impossible and then somebody broke it. And then all of a sudden, all these people broke the four minute mile. And it was just this incredible moment of human achievement. - Same thing with high jumps, right? 00:11:07,990 Like, forever people were like going over like front first and then some dude was like, no, I'm gonna go over backwards. And then they had the step-change, and just the amount of height you could get by high jumping. So love it, love it. I wanna get to the crux of what your insight was, and talk a little bit about data, because oftentimes, I think, in the world, everybody rushes to AI, and then they forget, actually the foundational element is the data, as you noticed. And you've done such an incredible job making that more accessible to everyone. So and one of the, a couple of things are sort of top of mind for me, in terms of challenges in the data space.

So the first one is a question of data sparseness. So, some industries that have been digitized over the last 10, 20 years, lots of data flowing through, right? Financial industry, lots of, sort of digitalization there, but like there's other industries like food and ad, like hell if I know how many farmers out there are willing to spend time putting sensors in their fields to gather data. And so, what are your thoughts in sort of making sure that we're bringing on the industries that may not be as digitized? How do we capture or help those industries capture more data, or maybe find ways to use less data in order to, you sort of build algorithms, or other ways that you've thought of? - Yeah, totally. 00:12:23,972 Yeah, and I think, this is actually one of the fundamental problems that we hope to solve. And I know like even without Scale it's going to be solved over the course of the next let's call it decade, because fundamentally you're exactly right. Like this, the way I think about this is, that there's sort of like these two waves of technology that are all sort of like crashing through the world so to speak. The first wave is sort of the internet, and more and more computing and more and more sensors. And that's great, because if you have more sensors, and you have more, the internet is more distributed, you have better coordination. Data gets saved more easily. You have automated workflows.

Like the software is amazing for a bunch of reasons. And then like, this interesting side effect is that, as that's happening, all these pools of digital data get collected in all these interesting industries, and all these interesting areas and verticals. And then those pools of digital data are actually, will create the opportunity for machine learning systems or AI systems to then have huge amounts of impact. And the sort of the second wave is, is AI and machine learning. That's going to build off the shoulders of the giants of like all this technology and sensors that have been deployed in the past. But then you enter a problem, which is that like, so if you think about the sort of quote unquote, old use cases of AI, like large-scale ad systems or a search, or ranking for social media, those use cases are very special, because they have what are called natural labels. Just by us clicking on things, on Google, or just by us clicking on a video in our YouTube feed, or clicking on something in our Facebook feed, we're providing labels about what are the things that interest us, what are the things that we're going to read, or spend time on or look at. And the reality is in almost every other industry, in every other part of the economy, you don't have that. You don't have these like natural labels that are just automatically, feeding into these machine learning systems, but you have lots and lots of data. And so most industries and most companies, whether that be in the healthcare space, or in financial services, or agriculture like you had mentioned, agriculture and food, or in sort of climate, there's huge amounts of data.

These companies actually have lots of lots of data, but it isn't, what we like to say, it isn't AI ready It's not annotated. It's not clean. It's not ready for the process of machine learning. And so that's sort of, in many ways it's like step one of what we've built at Scales. Like how do we enable every business to go from huge pools of unstructured data, towards well labeled, de biased, great, high-quality data sets, that are gonna allow them to build great machine learning systems of the future. And they think that like realistically speaking, that's gonna take some time, and there's lots of industries. But that's like this sort of like, that's almost this like critical step, or this critical bottleneck for AI to be deployed in significantly more areas. And the example I'll pose here, or that the example we talked about a lot internally is if you think about ImageNet, which was sort of the first, really large-scale, like dataset, computer vision dataset, that Fei-Fei Li's lab at Stanford helped create. That was the thing that then led to AlexNet and convolutional neural networks, and all these deep learning approaches, us realizing that they actually could be really effective and they could absorb all this data. And so there's this really interesting thing about how data will lead to, just having data alone, is going to lead to really incredible machine learning and great techniques being used.

And that's sort of one of the things that we get really excited about. The other thing that I'll mention is, there are plenty of use cases as well, where maybe it's difficult to collect data, whether that'd be for privacy reasons or that's for lack of sensors or whatever reason. And these use cases do exist, and sometimes they're incredibly important. And that's where sort of approaches around synthetic data or data augmentation, become really important, which is how do you take what little data

that you have, and really supercharge that, to be able to make it really, really valuable for building machine learning systems. And that's really something that we're very passionate about at Scale as well, it's like, how do we use synthetic data, data augmentation, to enable machine learning teams, to combine the benefits of real-world data with diverse and realistic synthetic data, to ultimately achieve sort of, really great machine learning algorithms. So it's kind of a mouthful, but data, data, data, lots of exciting stuff. - Yes, incredible. 00:17:17,370 For some of the students who may not necessarily be CS majors here. I do wanna talk about what quality data means and what data annotation means. And it's so fascinating to me.

Last week, we had a speaker come in, and talk about data in agriculture, and how before artificial intelligence, there must be human intelligence. So HI before AI. And she's really built her company at Gro Intelligence by hiring incredible experts who can interpret the data, and really connect the dots appropriately. So with your work, maybe I could also point out a personal example that I'm fascinated by. I do a lot of work annotating food data, and the chefs can tell the difference between a cilantro stem and a parsley stem. And I'm like, I have no idea. They look exactly the same to me. How do you work with your taskers, and how do you train them to annotate huge amounts of data in partnership with your customers? Because they're not necessarily experts in transportation, or experts in, like melanoma is on skin, like how do you then sort of help them train up so that they can do the work to ground truth. - Yeah, yeah. 00:18:28,539 And so a few things here, one of the fundamental beliefs that we have at Scale is that there's sort of like, there's sort of like a few resources that go into producing really high quality, machine learning systems.

If you think about like, what are the wrong ingredients? One of those wrong readings is compute. One of those wrong ingredients is data. And then the last wrong ingredient is sort of like human insight, so to speak, or like, and sort of the process of producing annotated data is combining raw data and human insight together, and sort of mixing those ingredients together. And all AI systems of the future are gonna be heavily reliant on human insight. Even if we have incredible advancements in the technology, et cetera, because at the end of the day, at minimum, we're gonna need oversight in these AI systems to make sure that they're built, they're producing the results that we would expect, and producing the results, that we think make sense and whatnot. And so I think, we think, this like problem, this sort of like, one of the core problems of machine learning is like, how do you make sure that for every problem, that you might have, that you might wanna solve using machine learning, you might wanna have a machine learning system do, are you able to effectively get human insight, to sort of like power that use case? And the way we think about it is like, is twofold. I think first is for folks who are trained up, you want those individuals to have as much leverage as possible. So you want, and this is somewhat circular, but it becomes very important. You want machine learning systems to do as much of the work as possible. And then really have humans spend all of their effort, on almost kind of the edges, or on the really like high judgment work that's required, that enables the upmost quality, while making them most efficient in their roles.

So that's the first piece. And the second piece is sort of this, this general education problem around, how do you enable people to be experts, and how do you like train them most efficiently? And one of the almost funny side effects of solving this problem at Scale is that I think we have one of the, probably the more interesting ed tech systems out there, in which we have systems internally, which track all the different kinds of, edge cases might be one way to talk about, but all the different kinds of nuances to some of these data problems, right? Like this cilantro versus parsley example could be one of them, and the different ways in which you can tell, or detecting a melanoma versus other kinds of growths on skin. So there's all these different nuances and we track all those, all these different nuances within our internal systems. And we are constantly trying to understand, for each of the annotators, which nuances or which edge cases are they performing well at, and they understand super well. And which ones might be tripping them up, and then proactively serving them materials and content, and examples that help them elucidate these cases that they might not understand super well. And so it's this, I really think about as an ed tech system which enables, enables people to really like, quickly grok and understand all the nuances of the data, and humans are incredible pattern recognition, right? If you think about, what it takes to go through medical school, and you talk to folks who have gone through medical school, it's an incredibly long process of just sort of like continued pattern recognition. And that's sort of, that same process where we try to distill for the annotators in our system to enable them to do great work. - That's incredible. 00:22:18,440 So first and foremost, it's interesting that you've actually, it sounds like you build AI to then help your annotators get better faster, which is recursive, right? You're actually using your own product on yourself to make yourself better over time. And a similar vein, I think, for a lot of students, AI can help make AI better, right? So oftentimes we talked about, you sort of talk about it.

We want the humans to pay attention to the thing that actually matters on an image, or in a sentence, and AI actually can be used to segment or pull out the part that requires attention. And so, again, there's all these recursive loops in here that I find totally fascinating with your work. Let's maybe broaden out a little bit. We had just kind of touched upon the melanoma example, and it's something that I find fascinating, but it's sort of a lot of sort of AI systems that are looking at skin for example, are focused on what's available already, right? So algorithms currently, as I understand it, are much more accurate for fair-skinned individuals because there's just more data. So when you think about like helping your customers, that you call your partners, sort of build, sort of fair balanced, sort of AI systems, how do you have that conversation with them? I know it matters less with roads and maybe like license plates where it's pretty unambiguous when you look at numbers, right? But like, how do you have that conversation with a partner who, when they're solving a problem, to aim for being responsible? - Yeah, definitely. 00:23:57,210 Well, I think first off to your point, I think responsible AI, is incredibly, for it to have responsible AI it's incredibly important to build high quality and representative data. I actually think it's critical. Because at its core, this thing that I talked about earlier, is that the data is almost like the food that you feed to these machine learning

systems and ultimately you are what you eat. And so the sort of, the data is sort of one of the fundamental areas where these biases or these kinds of really critical real-world issues can sort of stem from. And really the way we think of it, I was like, how do we build technology? How do we build tools? How do we build, how do we build things to enable either our customers or ourselves to ensure minimal bias and minimal harmful results with these machine learn, with the data sets that power these machine learning systems.

And or how do we build systems that are gonna flag that, or identify it when those biases may exist, as proactively as possible. So that then we know that we have to go fix that problem long before it ends up in the hands of a consumer, or some sort of critical decision making progress. So for example, we recently worked with like a bunch of medical researchers on automated medical imaging analysis. This sort of exact problem. We worked with the MIT media lab, and analyzed this exact problem you mentioned, of tons and tons of clinical images, and found there were a lot more light-skinned images than dark-skinned images. And what we did is, we actually basically help them replenish, or sort of de-bias the dataset, and add more data to the unrepresentative classes, either with real-world data, or using data augmentation, or synthetic data. And we're able to significantly de-bias the output, or the outcomes from the machine learning algorithm. And so I think it's not an easy problem in any way. This is actually a sort of like, this is one of the quote unquote hard problems of AI, in the sense that there's no easy solution. It's not like, I don't expect that next year we're going to get some, crazy machine learning architecture, that all of a sudden solves the problem of data bias, or machine learning bias.

It really is all about the nitty-gritty details about what is your data set look like? How do you know ahead of time, when there's bias in the dataset? And then how do you fix that? Or how do you proactively resolve that? And how do you keep going through that process? - In some ways that was a trick question, 00:26:43,810 because in order to unbias, you have to be aware of the bias. And so that is not always easy. So I know in the backend, you can unbiased once you're aware of the bias. Do you work with your partners to gather more data in the case where your sort of initial preliminary analysis is that there may be bias, do you go back to them and say, hey, you might wanna collect more in order for it to reshape this. - Yeah, we do. 00:27:13,260 And in fact, a lot of times, and going back to this like, a lot of times our customers, they actually, are sitting on incredibly large troves of data, but they don't have any of the tooling or infrastructure to help them sort of sift through the noise, and identify the quote unquote needles in the haystack, or identify the particular pieces of data that are going to drive forward their model performance, or drive forward bias the most effectively. And so we built tooling, our Nucleus product, which actually enables this exact process of like, hey, I know there's a problem with, let's say, not enough of a particular kind, of skin growth in my dataset. And then how can I then go through all my unlabeled data, sort of all my entire large troves of data, to just get more samples of this kind of skin growth, so that my algorithms are performing seemingly better in the future. And that's really the sort of like, much more so than even getting the data, it's about sort of instilling this sort of philosophy around continual improvement of algorithms, through continual improvement of data. And I think that that, if we can get there, then I think we're gonna be able to cope with all these challenges with bias, and data issues, and all that stuff, model issues, much more effectively.

- Uh, that's very cool. 00:28:34,833 Very cool. So maybe I could sort of zoom forward. We were starting from when you were 19, when you came up with Scale AI. If we were to sort of look out to 2032, I know you've worked in many different industries, healthcare, transportation, real estate, and more, what are you excited about? Like where do you think we're gonna be in 2032? And, with that, the secondary question is then, given that we have a lot of budding entrepreneurs in our class, and beyond, who might be watching on YouTube live right now, whether it's going deeper into picks and shovels as you've done with Scale AI, or other places within the ML gold rush that are like really, really interesting, like looking out to 2032, where do you think we're gonna end up, and what are the opportunities? - Yeah, no, I think it's, I mean, 00:29:27,033 I truly think, we're kind of in the golden age of AI, so to speak, where the sort of proliferation of use cases is gonna be absolutely massive. Like even at Scale, we started in autonomous vehicles, and the first few years of the company, it actually, it felt like machine learning felt kind of, I don't know if the word is lonely, but it was like autonomous vehicles were the big use case. Everything else almost felt like, a sort of side project, or something much smaller. And then fast forward to now, we see all sorts of like really exciting use cases in every single industry. So with financial services customers like Brex or PayPal, or Square or whatnot, we see interesting use cases around trying to understand, build systems that move money more effectively, or identify or understand transaction flows a lot better, or identify fraud much better. Or we see use cases with Flexport, which is sort of a global trade platform, and enable just like an incredible amount of efficiency in the process of global trade, which is very important.

It's really important that like we're able to get goods delivered from everywhere in the world, and it's an incredibly manual process today. And by using machine learning, you can automate a huge number of those workflows, and enable the overall economy to just get a lot more efficient, or whether it's with, whether it's with a large-scale autonomous, or automotive company, and a car company, and building not only full-stack autonomous vehicles, but also driver assistance systems, or systems that make drivers more safe, et cetera. So I think that like we're in this phase, where in this massive proliferation of machine learning systems. And I think that like the way, one way I would think about it is, the sort of like software eats the world mindset, is that you take, like, think about any industry in any, or any sort of like, any problem in the world today. And just imagine, okay, if you had software, how could you transform that? And I think that we've just seen, this has been this like very longterm, sort of slow transformation, because it turns out, humans are sort of like infinitely creative. And you'll take any system, and will be able to identify, oh, you can use software in this way, or, oh, you can use software in this way. And then you even, sometimes you even replace existing software with new software. There's like old enterprise systems, that are

replaced by like new style, more consumery kind of internet platforms. And so, I think it is gonna be this continual process where we're gonna look at something, we're gonna look at a problem. Let's say, like in insurance, the process by which claims get processed.

We're gonna look at the problem. And we're gonna think, okay, if you act like, if you use machine learning the right way, and not just AI in some magical sense, but actual, like the core fundamentals of machine learning, then you can design this process to be 10X more efficient. And we're just gonna keep identifying all those problems. And I think you go fast forward to 2032, it's gonna be everywhere. And the opportunity won't have stopped. Like we're still gonna have like plenty of opportunity, to apply AI to these systems. I think maybe more, to just name a few specific examples, 'cause I think these are maybe some of the ones that are quote unquote cooler, or more exciting right now. I think that there's a few that I think are really important. And for those of you with a course, or like thinking about what ideas are exciting, I think these are maybe some areas to think about. I think first one is science.

Science is, there's actually these papers about how scientific progress, has kind of been slowing actually a little bit over the past few decades. And one part of that is that like, when you think about science, let's say a century ago, or two centuries ago, you could do so many experiments, and your ability to validate your ideas was really, really exciting. And now we're at a point where like a lot of the cheap experiments, so to speak, have been explored. And now we have like ridiculously expensive experiments. like particle accelerators are extremely, extremely expensive. Or large-scale clinical trials are very, very expensive, et cetera. And one of the really exciting use cases of AI is using AI to simulate, effectively, basically, simulate experiments, significantly more effectively than you could in the past, or that you could use in classical methods. And there's already a lot of examples of this in whether that's something like, in AlphaFold out of DeepMind, or using AI applied to sort of like fusion experiments, or fusion simulations. But I think there's gonna be a huge boon in physics, chemistry, biology, pharmaceuticals. And it's going to transform a lot of the sort of like, it's gonna be this base technology that empowers a lot of future innovation.

So I think that one's really critical, Metaverse is a use case, is something that a lot of people are talking about these days. I think it means a lot of different things to different people. But I think if you think about AR or augmented reality, which is probably one of the form factors that might, that probably feels most intuitive to a lot of us, which is that, hey, we're just gonna have sort of this digital overlay over sort of our natural lives. If you think about that problem, it is an incredibly complex machine learning problem an incredibly complex AI problem, because fundamentally, you need to understand the world. And how these different objects relate to one another. And how I, as a person, can relate to those objects. And if people are walking past each other, and they'd make a look, you need to be able to understand that kind of stuff. And so you need to have this very fine grain understanding of what is going on in the world around you. And that's a very, very challenging AI problem. But I think that it's one that, is going to enable these sort of like, very sci-fi like consumer experiences, that I think, of the future, that I think, we're all really fundamentally, really excited about.

And the last one that I'll kind of mentioned, just cause, I think, this one is really important. And it's somewhat controversial. But I think it's an important one to talk about, is, I think AI as applied to the government's problems, in particular applied to sort of national security, defense intelligence, et cetera. And I think that, we're in a very interesting period in the world where the sort of like, warfare's shifting from sort of a previous paradigm to a significantly more digital paradigm. And now, a significant number of these sort of like skirmishes of the future are going to happen entirely digitally in cyberspace, or via AI systems or via purely digital systems. And I think it's really, really important that if you believe in kind of democracy and you believe in the values that the United States represents, that the United States and other democratic countries are able to utilize best-in-class technologies to not be vulnerable, as a sort of long-term platform shift is occurring. And so I think it is really important that we have some of the best and most brilliant technical minds thinking about how do we build the best-in-class systems for the future of the United States, to enable the United States to be as effective from a defense and sort of intelligence perspective, as it has been for the past, call it 50 years, which has really enabled the sort of like, modern era of peace. So I think those are some of the areas, that excite me the most, slash I think are most important. - Uh, incredible. 00:37:31,300 So just to sum it up, science, Metaverse, AR, VR, and then government.

It's fascinating to me, if I were to tackle that question, I'm almost like, what are the chores that human beings don't like doing like washing dishes or like doing my taxes, right? Like one day, all the things that are taking up our time and it could be automated. I see 2030 as a point in time where all those things are automated and we are free to actually do what only uniquely humans can do, which is to spend our time being creators and being creative. And I hope, that's the future that I look for. And I think I would support the comments that you make here. Let's talk about robots for a second. At one point, (laughs) in early days of some of the everyday robots work at Google, I would have a robot come to my desk and try very hard to clean up my desk. And it would fail. Generally, would try to pick up my bowl of cereal and accidentally pour milk all over my desk, or like once in a while it would like steal my bottle of wine, like old stuff like that. Putting that aside, I'm actually a little bit curious about with AI and data, and so tied with the conversation that we had around AR. In the future, there's going to be artificial humans, right? And I heard an interesting podcast recently with Ezra Klein interviewing Ted Chiang, who's a science fiction writer.

And he said that before machines become sentient, we will probably end up causing them to suffer a whole lot. And this kind of begs the question. AI, how do we give feelings to machines, and are machines gonna have feelings? And is that a

bridge too far? And I know were sort of steering away from sort of data that the whole entire purpose of the work that you're doing is to really create this foundation for data, to then bring AI into the future, to make human life better. Curious how you feel about robots and machines in that space. - Yeah, no, 00:39:46,250 I think it's a super interesting philosophical question, which is that, like, let's take a- Right now in a Stanford AI class, most of the students will train some sort of simple neural network to recognize objects and imagery. Is that neural network, is that like, does that neural network have feelings? Like when you will delete it from your hard drive because you're running out of space. Is that? Does that cause suffering to that neural network? And I think that, I think there's a super, it's like very tough questions to answer. Like, where are the boundaries? What are the? For each of these edge cases, what does that kind of look like? I actually think that, I think the sci-fi fantasy states in the future where you have AI systems that actually fully resemble humans in terms of their ability to have judgment, and communicate with us, I think it's actually quite far. Even my evidence point is like, we're still not at a point where an AI system can fully accurately process documents, and OCR are like still, like, does not work fully effectively. And the judgment that goes, the judgment that will be required from an AI system to actually like resembles people is so far off that I think that, I think we're gonna be, in this sort of like limbo state, so to speak, where we have AI systems that are very effective, and very useful.

And can do things that like, maybe we didn't think they could do like three years ago, but you're not gonna get to the point where you have AI systems that are, that you would sort of like, where there's like these really deep philosophic questions that are tough. One thing that I do think though, is going to happen probably sooner rather than later. And this is maybe the sort of like near-term problem is that we're going up more and more, have AI systems that affect how humans think, right? And this is sort of like, it's already happened to some degree where social media systems or ad systems, or even search, these are machine learning systems. They pick content for you, and that the way they pick content for you will change how you think. And this is only gonna keep going in that direction. Like, I think in the future, we're going, like not so distant future, we're gonna have, children are gonna have sort of like AI friends. And they're gonna have these chat bots that they can talk with, and practice speech with, and sort of like, and learn from, and talk about their days, and like basically like true friends. And they're gonna affect how humans think, just like how phones have sort of like changed the human's attention span, probably permanently. And so I think that those, there are sort of these like interesting social issues around what does it mean for AI to have such a real tangible impact on how we might behave in time? That I think are, I mean, those are like, that's a really important sociological question and what are the implications of that? How do we get ahead of it? And I think ultimately, at least from our neck of the woods, in our view, I think a lot of that comes down to, again, how do we ensure accuracy, quality, and efficacy of the data that are feeding these ML systems. So we never end up in like, a really bad scenario where the data distribution, or the data used to feed in one of these algorithms, was so off base that it caused all these unintended consequences, sort of the people that were using them.

And in some sense, I think that, like, if you think about, sort of like political polarization, that's happening. On some level, that is a data distribution problem. It's like, hey, what's happening is that these ML content systems are favoring more and more polarized content. And the data distribution has shifted so significantly, and that's causing all these like weird social dilemmas that are really tough to think about and tough to deal with. And so anyway, long story short, I think it does all come back to data, at least like, I think data is a necessary, but not sufficient, part of the problem. And I think that AI is gonna change how we think pretty soon. And we're gonna have to think about that. (Emily laughs) - That was fantastic. 00:44:27,370 Thank you for being so just genuine in responding to a pretty provocative question that I was asking you. So I know that the students are hankering to ask you questions.

So let's dive into that, for the final remainder of this time. First question here. What are some key challenges you have had to overcome while running your company? - Yeah, yeah, yeah, no. 00:44:52,950 So it's a great question. And it's almost, answering is almost impossible because it's almost like, what challenges have I not had to overcome in building a company? I think that one of the things about, I think it's probably like doing anything challenging and sort of this concept of like, whether it's building a new product, or building a company, or building a team, there's just, it's like whatever, like you'll deal with so many challenging scenarios in time. And there's sort of like this infinite chaos to the world that will cause you to sort of be confronted with the most strange circumstances that you like could not foresee. And you're just gonna have to figure them out. But I think that, like, if I had to distill it down, I think, probably the parts about building Scale that are really maybe the most challenging, but also most rewarding, have been about building the team, getting great people to sort of like join the mission, and really see the potential of what we're able to accomplish, whether that be people who I work with, or the incredible team at Scale, whether it be our amazing investors, whether it be our amazing customers. But I think this sort of like, this fundamental challenge of getting people to buy into what you believe the future to look like, I think is one of the hard problems of the world. And I think, I felt very lucky, that so many people have chosen to take a chance on me.

- You have an incredible vision and very, 00:46:30,670 a lot of clarity around that vision. So I can understand why a lot of people have joined you in whatever way they can. I think that's also underestimated. I think a lot of our engineering students here, like, oh, I'm gonna spend a lot of time on the tech, but it turns out that spending time on people is just as important if not even more important. All right. Next question. At what point during the development of your company, did you realize that number one, it could become a serious competitor in the AI space, and number two, that you were ready to pursue building at full time? - Yeah, no, it's a really good question. 00:47:08,693 I think that early on sort of, it always is helpful to be wildly optimistic, I should say. Or at least dare, maybe the proper terminology is to dare to dream big, which is that even early on, I think this sort of like this thread of thought, which was, hey, if we actually, do solve this sort of like data problem, for the

machine learning community, that is big. And I don't know exactly, what it's gonna look like to build from that point, but we're gonna be able to do a lot of interesting stuff if we get there.

And so that thread of idea did emerge, and it didn't become, it honestly wasn't relevant for us for the first many years of the company, 'cause we're just focused on data labeling, data annotation, we're just focused on like solving that. But then you sort of, we solved that problem. We're really focused and we got Scale. And then we picked our heads up and we're like, wow, we're actually at this sort of like the promised land that we thought at the very beginning. That like, we actually have a shot at like, building a lot more, doing a lot more, more industries, to sort of accelerate AI and machine learning development. And so, I would say that it was a, maybe early on, it was a pipe dream. And then it was sort of like many years later before it actually felt like, hey, this is like a real opportunity for us. And the second part of the question, which was at what point did I decide I was ready to pursue it, building at full-time. I actually, like, I wish I could say, oh, I was just like so convicted. And so I just, and so I just like went for it, but actually it was, I really started working on it in earnest when school ended.

I remember even like I put working on all of it on hold. What, to get through like my finals, and my final projects and all that stuff. And then the summer started and I was like, all right, I have nothing else to do. And then I was, we were very fortunate that we got, over the course of the summer months, enough momentum, where I was like, hey, there's kind of no option but to just keep going at this thing. And so I think it's, I think it's sort of like this barrier, this mental barrier feels very large, oftentimes around like, hey, I have a life right now. At what point am I gonna be ready to like sacrifice my life today and then go do, and then go approach this other life. I think the, and I think a lot of times it'll seem very irrational. And maybe my advice is to like, not actually be sort of quote unquote, trapped in that false dichotomy, but actually just put yourself in situations where you have sort of freedom and flexibility to explore and then like, see how you gain momentum in that scenario. - I appreciate that, 00:49:54,330 because sometimes students feel like they have to have something lined up for the summer, but it was actually a gift for you not to have had something lined up and to have that space to kind of really dive deeper into something that you really cared about. Similarly, let's take the next question here, knowing what you may know now, after running your company for three years, if you could go back to the first day you started your company, would you do anything differently? - Yeah, well, there's always a butterfly effect 00:50:18,860 where like maybe if I did something slightly differently then like, that all these other things would have gone totally differently.

And so on the whole, I wouldn't trade sort of, I wouldn't screw it up too much. That being said, I do actually think that like, the biggest thing that I would say, like if I were to go back in order to tell myself one thing, it would be that, like, it's the people matter so much. And like however much, I was really focusing on people early on. And a lot of that was guided because of mentorship that I had gotten or advice that I had gotten. And also just the sort of like this like selfish desire. I was like, hey, I wanna work with people who like really inspire me and really get me excited. And that seems like the best way to live, especially I'm gonna be working on this all day at night. And so that was like, I ended up getting there a little bit. But I would just, it's like now, I know in my mind state that it's like, however much, I thought that people mattered at the start, they matter like 10X. It's like obviously they matter 10X more now.

And I wish you could just sort of like, transport that thought back to my prior self. So that's kind of tough. I also think maybe the other piece is like, along the way there's a lot of moments of like great self doubt. And that's just, it's almost like part of the journey, 'cause almost invariably, you're just gonna have a situation where it's like, wow, crap, I've got almost like no idea what to do. And I think probably the big thing I would tell myself is like, it is like totally okay to be in that spot. It is totally okay for that to happen to you. It is not okay to be paralyzed when that happens. And even if you literally, take the worst possible action coming out of that scenario, it'll be better to have taken that action, than to have taken no action. And so it's maybe, that's the other thing I would tell myself. - Ultimately, there are no, 00:52:23,940 I would say, one way doors, right? I think by taking an action, we can always course correct along the way, but if we don't take any action, we don't have any new information.

So it's super important to actually make a decision act, even with imperfect information. All right, next question. With recent scrutiny regarding data privacy, how do you think AI companies, who are of course relying on good data, should go about data collection, huh? - Yeah, no, I think this is a super, 00:52:51,369 it's a super important question. 'Cause I think that the question is like, and this has spurred in, there's been a bunch of examples in the past that have spurred this. But some of them, for example, are like, the realization that Siri or Alexa. are learning from recordings, and you had no idea that they're necessarily learning from those recordings. And so, I think it's like super important, almost like reorientation that the entire technology or ML community needs to take, which is like going from, going from having it starts like, however, I can get the data, I will use that data, and then train machine learning systems on it, to, I need to get this data in a very responsible way. I need to do it in a way that is like, where everybody who's providing the data understands that their data is being provided for these AI systems, where I'm collecting it in a de biased way, where I'm collecting sufficient amounts of it, et cetera. And I think that, the answer is that that's gonna make, that's going to make it harder to build the same kinds of AI systems, potentially, but it's the way that technology needs to develop. Like it's not a tenable state, for machine learning systems we developed sort of, by taking advantage maybe like, a checkbox that nobody reads or stuff like that.

And so, I think, it's a very important shift. It's a big shift that's gonna happen. And we're helping a lot of our customers think through that, and make them robust to ensuring that they actually get data for these ML systems that is sort of responsibly collected. - It actually goes back 00:54:35,740 to something you've said many times, even publicly, like, your

customers, you partner with them, right? You're not just providing them with a solution. You're really truly partnering with them, and being in dialogue with them, because these challenges like privacy are very nuanced and tricky and you can't, it can't be the same thing every time, right? And so you have to sort of lean in, and have the conversation as a partner, as opposed to selling a thing that's finished. And so I'm really glad that that question was asked 'cause it is something that's top of mind for so many people. Let's tackle one more question, one more question. Try to fit it in there. What have been your most significant failures? Oh, goodness. And what did you learn from them? Or like, let's start with just one big failure that you felt was a failure, and how did you learn from it? - Yeah, I think.

00:55:20,680 Man, there's so many. But I think, (laughs) I'll talk about one that is very much top of mind. We had this one, one of our first customers, and really our first very large customer. And I won't name them, but they're sort of like a, they're a prominent company. And they really, in the very early days of Scale, they took a huge chance by working with us. 'Cause like, obviously, our product was really janky. It only kind of worked. They took a huge bet on us that we would be able to solve this problem for them. And actually for a while, we had a pretty successful relationship with this customer and that's what allowed them to grow to be very large. And will allow us to scale up with them, and for us to sort of like, to build this great relationship with them.

And then one of the, I think it was like, I don't know if it was a, I'd call it a failure or a mistake or what not, is that we ended up losing focus on that customer. And it was sort of like, okay, let's go find another big customer. Let's go do all this other stuff. And in doing that, we're in that process, we ended up, we ended up really like in the ignorance, in ignoring that customer, they, at a certain point, we just, we weren't actually solving their problem. We weren't actually sort of like, being the scrappy team that they could bet on, they could actually like bet on to build an incredible product for them. And they ended up leaving. And I remember this was like this incredibly, it was very challenging in the moment. This is one of the examples of like, when you talked about, hey, you're gonna have moments of great self doubt. And I remember asking myself like, hey, is this, do we even have a business? Like, does it even make sense for customers to work with us? Like, asking myself all these questions? And it was a really tough moment 'cause it was also one of these things where I was like, there was almost nothing that we could have done to convince them to have continued to work with us. So he was like, this was his warning.

So he was like, what's done is done. And the big takeaway, or the big thing that we learned from it is like, it's just so, so important to be customer centric. And it's so important to always be focused on your customers and make sure that you're always understanding what you need to do, and how you make them just ecstatic about your product, and how you sort of like exceeded all their expectations. And that moment really like, however embedded this concept was in our culture beforehand, it just like, it just made it even more visceral, and made it more obvious. And so I think that was an incredible learning moment. The only thing is like, I'm very grateful that it came so early on in our journey. (gentle music)..