[Will AI Save Humanity or End It? with Mustafa Suleyman Trevor Noah]

Introduction

So I feel like people fall in one of two camps on AI. They either think it's going to destroy all of humanity. 42% of CEOs surveyed fear artificial intelligence could destroy humanity. This is something that put in the wrong hands, could destroy humanity, or they think it's going to solve every single problem. Suleyman Mustafa Suleyman Mustafa Suleyman is an artificial intelligence pioneer. He is the AI CEO at Microsoft. He is very big in the artificial intelligence world. How do we manage these technologies so that we can coexist with them safely? Can humans and AI coexist with each other peacefully without one taking over the other? This is What Now with Trevor Noah. Mustafa. How are you, man? I'm very good man. This is great. This is good. It's funny that there's this. There's almost 2 or 3 different types of conversations I have with people. There's ones where I'm hanging out. It's my friends we're discussing. Just whatever, you know, shooting the shit. Then there's some way I bring a person in and I'm trying to, like, get something from them or learn about their world. And then there's the third type of interview that often stresses me the most, because I feel like I'm speaking to a person who is who has like an outsized influence in our world. And if I mess it up, I don't ask the questions that the world needs. And I feel like you're one of those people, because even before your current job, you were considered one

Early DeepMind

of like, if there was a mount Rushmore of the founders of AI with none of the with none of the baggage, with none of the baggage of modern colonial history, no colonial history. One yeah. No colonial history. But if there was, like, a large Mount Rushmore. Your face would be up there, you know, as being part of DeepMind and the founders of DeepMind. And then now you are helping Microsoft, like one of the biggest, you know, tech companies in the world by, you know, market cap and just by influence, shape its view on AI. And so maybe, maybe that's why I wanted to start this conversation because it's it almost feels like where we meet you in the journey. Now, you know, what would you say has been the biggest shift in, in your life and what you've been doing in AI, going from a startup that was on the cusp and this fledgling world of AI, to now being at the head of what's going to shape all of our lives in AI. Wow, what an opener. I mean, seriously, no pressure. You know, the crazy thing is that I've just got incredibly lucky. I mean, I was strange enough to start working on something that everybody thought was impossible, was totally sci fi. That was, you know, just dismissed by even the kind of best academics, let alone any of the big tech companies didn't take it seriously. I mean, 2010, you know, just to really ground on that, we had just got mobile phones three years earlier. Yeah, the App Store was just coming alive. You couldn't even easily upload a photo from your phone to, you know, an app or the cloud. Right. And somehow, you know, my courageous, visionary co-founders, Demis Hassabis and Shane Legg

How AI Works

had the foresight to know that, you know, technology, ultimately digital technologies ultimately become these learning algorithms, which, when fed more data and given more compute, have a very good chance of learning the structure and nature of the universe. And so I was just very privileged to be friends with them. Be part of that mission to you know, I was only 25 years old and, um, you know, kind of had the fearlessness to believe that if we could create something that truly understood us as humans, then that actually represents

one of the best chances we have of improving the human condition. And that's always been my motivation to create technologies that actually serve us and make the world a better place. Like I was cheesy before, it was cheesy pre cheese. You said something that like that sparked a thought in my mind, and I think a lot of people would love to better understand this. We see headlines all the time saying AI this AI in your job AI does that the thinking that not thinking though and you said something that engineers will gloss over quite quickly. You'll go data and compute and then a model and then it help help me break that down. Help me like just explain that to me in the simplest terms possible. What changed? And what are you actually doing? Because it's like we always had data, right? We've had documents, we've had files, we've had information. We always had computers. Well, not always, but we had computers for decades. What what changed? And what is AI actually coming from? So the the best intuition that I have for it is that our physical world can be converted into an information world. And information is basically this abstract idea, like it's mathematics, it doesn't exist in the physical world, but it's a representation of the physical objects and the algorithm. Sounds complicated, but really it's just a mechanism for learning the structure of information and the relationship of one pixel to another pixel, or one word to another word, or one bit in the audio stream to the next bit in the audio stream. So I know that sounds very like abstract, but the structure of reality or the structure of information is actually a highly learnable sort of function. Right? And that's what we saw in the very early part of AI between 2010 and sort of 2016. These models could learn to understand, or at least not understand, but maybe they could learn to generate a new image just by reading a million images. And that meant

Do Machines Think?

that it was learning that, you know, if an eye was here and another I was there, then most likely there would be some form of a nose here. And although it didn't apply the word nose and I just had a statistical correlation between those three objects, such that if you then wanted to imagine, well, where would the mouth be? It wouldn't put the mouth in the forehead. It would put the mouth just below the nose. And that's what I mean about the structure of information. The algorithm learns the relationship between the common objects in the training data, and it did that well enough that it could generate new examples of the training data. First, in 2011, it was handwritten black and white digits. Then by 2013 it was like cats in YouTube images, which was what Google did back in 2013. Then as it got better, it could do it with audio and then over time, you know, roll forward another ten years. It did it with text. And so it's just the same core mechanism for learning the structure of information that has scaled all the way through. So it's interesting I heard you say what it does. And I also noticed at a moment you said it understands. And then you said, yeah, well no wait. And I've actually noticed quite a few engineers and people who work in AI and tech struggle with explaining it to layman's using, you know, human language, but then very quickly going like no, no, no, no, no, it's it's not human language. It's like, does it think or does it do what we think thinking is? Yeah, I mean, this is a profound question. And basically it shows us the limitations of our own vocabulary because what is thinking, you know, that it sounds like a silly question, but it's actually a very profound question. What is understanding? If I can simulate understanding so perfectly that I can't distinguish between what was generated by the simulation and what what was generated by the thinking or understanding human being, then if those two outputs are equivalently impressive, does it matter what's actually happening under the hood, whether it's thinking or understanding or whether it's conscious. You know, it's a very, very difficult thing to ask because we're kind of behaviorists in the

sense that as humans, we trust each other, we learn from each other, and we connect to each other socially by observing our

Humanist AI

actions. You know, I don't know what's happening inside your brain, behind your eyes, inside your heart, in your soul. But I certainly hear the voice that you give me, the words that you say. I watch the actions that you do, and I observe those behaviors. And the really difficult thing about the moment that we're entering with this new AI agent era, as they become not just pattern recognition systems, but whole agents, is that we have to engage with their behaviors increasingly as though they're like sort of digital people. And this is a threshold transformation in the history of our species because they're not tools. They're clearly not humans. They're not part of nature. They're kind of a fourth relation, a fourth emergent kind of. I don't know how to describe it other than a fourth relation. Yeah. I mean, you've called AI the most powerful general purpose technology that we've ever invented. And when I when I read that line in your book, I was thinking to myself, I was like, man, you are now at the epicenter of helping Microsoft shape this at scale, you know? And and then it made me wonder, what are you actually then trying to build? Because everyone has a different answer to this question, I've realized. You know, if you if you ask Sam Altman ChatGPT Sam Altman says, I'm trying to build artificial general intelligence. And I go like, oh, I like the app. He's like, I don't care about the app. Actually. I want to make the God computer. And then you speak to somebody else and they say, oh, I'm trying to make AI that can help companies. I'm trying to make AI that helps. So what are you actually trying to build? I care about creating technologies that reduce human suffering. I want to create things that are truly aligned to human interests. I care about humanist superintelligence. And that means that at every single step, new inventions have to pass the

Philosophy vs. Tech

following test in aggregate. Net. Net. Does it actually improve human well-being, reduce human suffering, and overall make the world a better place? Yeah, and it seems like ridiculous that we would want to apply that test. Like, surely we would all take it for granted that no one would want to invent something that causes net harm? Yeah, right. But, you know, there's certainly been other inventions in the past that we could think of that, you know, arguably have delivered net harm. Right. And we have a choice about what we bring into the world. And so even though it's in the context of Microsoft, the most valuable company in the world today, we have to start with values and what we care about. And to me, a humanist superintelligence is one that always puts the human first and works in service of the human. And obviously there'll be a lot of debate and interpretation over the next few decades about what that means in practice, but I think it's the correct starting point. I've always wondered how the two sides of your brain sort of wrestle with each other around these topics because, you know, someone asked me, they were like, oh, who are you having? On I go? Mustafa is coming on Mustafa Suleiman. And they're like, oh, what is he doing? I explained it a little bit and I was like, oh, so like a like an AI guy. Then I was like, yeah, but he's also a philosopher and they're like, what do you mean he's a philosopher? I was like, no, no, no. Like actually like actually this is somebody who studied philosophy, is engaged like you, you, you, you think about the human ramifications of the non-human technologies that are being built by and for humans and and what you know, what it is for me is I always judge people by what they choose to. Yada yada, if that makes sense. You know, so I've talked to some people in tech and I say, what about what about the dangers? And they go, well, look, I mean, of course we've got to be aware of the dangers, but but the future

and it's so big. And then I remember once I met you the first time, actually, I met you. I said, uh, the technology is amazing. And then you went the dangers. Let me tell you about the dangers. Let me tell you about the things we need to consider. And I was like, what just happened here? Do you know what I mean? I was just like, what is this? Is this guy working against himself? And so I wonder now, like, when you're in that space, when you're working on something that is that big, how do you find the balance? Because we would be lying if we said humans could live in a world where we could ignore technologies. I've seen people say that my opinion is that you can't ignore a technology, right? You can't just be like, no, we'll act like it doesn't exist. But on the other hand, we also can't act like the technology is inevitable because then we've given ourselves up. So when you're the person who's actually at the epicenter of trying to build our future, and I know it's not you alone, please don't get me wrong, but how do you how do you think about that? How do you grapple with philosophy versus business, philosophy versus technology, human versus like an outcome? What are you thinking of? You've called out my split brain personality, and now I'm like, thinking which side of me should answer? I can answer twice. Yes. Twice. You can pick your answer. I'll give both. I think part of it is just being English. You know, I'm kind of like, I'm. I'm more comfortable than the average American thinking about the kind of cynical, dark side of things. Those rainy days. It's rainy days, man. It's those rainy days. And I just think, I don't know. Like truth exists Lists in the honesty of looking at all sides. And I think if you have a kind of bias one way or another, it just doesn't feel real to me. And I guess that's kind of my philosophy or kind of academic side. That is a core part of who I am. Like, I, I'm comfortable,

Future of Energy

you know, living in, you know, what to some people might seem like a set of contradictions, because to me, they're not contradictions. They're truth manifested in a single individual. But if you are honest about it, it's also manifested in every single one of us, too. Um, you know, I happen to be in a position, but like, the company has to wrestle with these things. Our governments have to wrestle with these things. Every single one of us, as citizens has to confront this reality. Because, you know, every single technology just accelerates massive transformation, which can deliver unbelievable benefits and also create side effects. And it's like that. That idea has been repeated so many times. It now kind of sounds trite, but once you get over the trite part, you still have to engage with the fact that the very same thing that is going to reduce the cost of production of energy over the next two decades by 100 x, reduce the cost of energy by 100 x. You think that I can do 100%? Like I feel very optimistic about that. So then I'll say that again. So reduce the cost of energy. I think energy is going to become a pretty much cheap and abundant resource. I mean, even solar panels alone are probably going to come down by another five x in the next ten years. Like just that breakthrough alone is going to reduce the price of most things. Um, and what is that? What is that through? Is that like the AI being more efficient, teaching us how to create different energy grids, teaching us how to create energy differently? Like what? What what would you predict it coming from? Well, I mean, so at the most abstract level, these are pattern matching systems that find more efficient ways than we are able to invent ourselves as humans for combining new materials. Now, that might be in grid management and distribution. It might be inventing new synthetic materials for, um, you know, batteries and storing renewable energy. It might be in more efficient solar, you know, voltaic cells that can actually capture more per square inch, for example. I mean, there are so many breakthroughs that, you know, we're kind of on the cusp of that require just 1 or 2 more

pushes to get them over the line, even the superconductors from last year. Those things, any one of those could come in, right? And if they do, we see

AI's Environmental Cost

massive transformation in the economy. I mean, imagine if by 2045, you know, energy is let's say 10 to 100 x cheaper. We will be able to desalinate water from the ocean, um, anywhere, which means that we would have clean water in places that might be 50 degrees or whatever, you know, 120 degrees hot. Right? Which means that we can grow crops in arid environments, which will mitigate the flow of migration because of climate change. Which means that we could run AC units in places that we never could before. You know, there are so many knock on effects of fundamental technologies, general purpose technologies like energy coming down by 10 to 100 x. Um, so there are huge reasons to be optimistic that everybody is going to get access to these technologies and the benefits of these technologies over the next couple of decades. Yeah. And that will make life much easier and much cheaper for everybody on the planet. So let's, um, let's jump into that a little bit. Like it could make energy how many times cheaper? Well, I was saying 100 x cheaper over 20 years, 100 X cheaper over 20 years. So this this is this is one of those instances that I've struggled with because you know, like depending on where you get information and how you get information, it changes how you perceive the issue, right. So I remember being really angry when I saw how much water is consumed by like typing one query into copilot ChatGPT any any AI model. Then I was even more angry when I saw how much water is consumed by like getting a picture, you know, made. And then I saw something else that was like, oh, this is nothing compared to cars and, you know, produce and like making hamburgers and that. And then I was like, okay, like, where's the information coming from? Where's it not coming from? The response to the price of, of of AI. Like is it is it driven by the AI industry saying, no, this is actually not that bad? Or like, how do you think we should look at it or how do you look at it? I mean, look, it consumes vast amounts of resources precious metals, electricity, water, no question about that, right? Um, on the energy side of things, all of the big companies now are almost entirely 100% renewable. Certainly Microsoft, 100% renewable. I think we have 33GW of renewable energy in our entire damn fleet of computation. Uh, for comparison, I think Seattle consumes two gigawatts per year of energy. So just to put that into perspective, the whole of Seattle, the whole of Seattle consumes two gigawatts. And Microsoft is creating how much? 33 overall in the fleet. But this is wild. Yeah. No no no. But still and the vast majority of it is 100% renewable. So coming from solar or wind or or water, but it also consumes a lot of water in the process. Like we have to cool these systems down and you know for sure that consumes a lot. Now, I don't know that there is a easy way of, you know, there's there's no shortcut there. It's expensive. It consumes, you know, resources,

Scale & Risk

consumes a lot of resources from the environment. Um, but I think net net, when you look at the beneficial impact to me is justified. I'd like. You know why you wouldn't give up your car or tell people to give up their car anytime soon because it uses aluminum and rubber. And is this an essential part of your existence? And I think AI is going to become an essential part of everybody's existence and justify the environmental costs, even though that doesn't mean that we have to go and consume diesel generators and, you know, carbon emitting, like we get to start again from scratch, which is to say, new technology arrives, new standard has to be applied to it, which means that our water has to also be cleaned and recycled, which, you know, many of the data centers do now, you know, take full lifecycle responsibility for

cleaning the water. And the same with the energy. It has to be renewable. So there's no easy way out. It's just a rough reality that producing things at this scale is definitely going to consume more resources from the environment. It's funny, every time I try and think of it, I, I think of the, um, the gift and curse that comes with anything that scales, Ills, you know, I. The analogy. I always use it for myself as I go. I think of like an airplane. Before an airplane is invented, especially like a large jumbo jet. The amount of people who can die while being transported is much lower. Really? If we're honest. You know, a car, four people, six people, whatever. It might be still tragic, but a smaller number. The

Jobs vs. Workers

plane comes along, you can go further, you can go faster. But it also means there can be something more devastating on the other side of that plane crashing or something going wrong. And it feels like that scales with with AI as well. It sounds like you're saying to me on the one side of it, this technology could completely change our relationship with economies and finance and society. But then there's there's the looming other side of it that that could crash. And so maybe that's a good place for us to start, like diving into this is what's noise and what's very real for you as somebody who sees it, because everyone gets a different headline about AI, it doesn't matter where you are in the world. It doesn't matter your religion, your race, whatever it is. Everyone gets a different headline about AI. But when you're looking at it as somebody who is working on creating it every single day, what is real and what is noise in and around AI? So I think it's pretty clear to me that we're going to face mass job displacement sometime in the next 20 years, because whilst these technologies are for the first part of their introduction, augmenting like they add to you as a human, they can save you time. Yeah, it's like a bionic leg, but for like cognitive laborers, you know, uh, I think like you could, you know, who is it? I think it was Steve Jobs that called it, like, the bicycle for the mind. You know, it's just sort of exercising, you know, digital technologies allow you to exercise new parts of your mind that you didn't know you had access to. And I think that's definitely true. But much of the work that people do today is quite routine and quite predictable. It's kind of mechanized. Um, yeah, like cognitive manual labor. And so that stuff, the machines are going to get very, very good at those things. And the benefits to introducing those technologies are going to be very clear for the company, for the shareholder, for the government, for, you know, and so we'll see, like rapid displacement and people have to figure out, okay, what what is my contribution to the labor market? I think those fears are very real. And that's where governments have to take a strong hand, because there needs to be a mechanism for taxation redistribution. Taxation is a tool for incentivizing certain types of technologies to be introduced in certain industries. And so it's not just about generating revenue, it's about limiting, adding friction to the introduction of certain technologies so that we can figure out how to create other opportunities for people as this transition takes place. Yeah, it's funny, one of one of my favorite quotes I ever heard was, um, I think it was Sweden's head of infectious diseases. I think that's what his job was. I spoke to him during the pandemic, and we were just talking about life in Sweden and what they do, and and I asked him a question about Labor and Sweden and how everything works out there. And and he said something fascinating. It was, um, no, I think he actually was in the Labor Department on that side. He said, in Sweden, unlike in America, he said, in Sweden we don't care about jobs, we care about the workers. And I remember that breaking my mind because I went, oh yeah, everyone always talks about like the job, as if the job is something that is affixed to a human being. But really, the human is the important part of the equation. Yeah, the job is just what the human does. And so our focus has to be on making

sure that the human always has a job. But from what you're saying, we don't know what the job will be, because the jobs that we know now are sort of easy to replace. Yeah, it's data entry, data capturing, sending an email, doing an Excel spreadsheet. That stuff is easy. Actually, when it comes to AI. And then now we don't know what the next part of it is. And so maybe my next question to you then is when you're in that world, the philosophers out of your brain, like, what do you what do you think the onus is on us and the like? I mean, the tech companies and all the to work on discovering what the new job is or do we not know what it will be? Well, but also I would tweak what you said that are the job of society or the function of society is to create jobs that are meaningful for people. Yeah, like I'm not sure I buy that. Like I think really many people do jobs which are super unfulfilling and that they would be quite happy to give up if they had an income. This is true. Like that's true. Like, we're probably very lucky that we get to we get paid for the thing that we would be doing if we didn't get paid right. I would certainly be doing that. And so I think the function of society is to create, um, a peaceful, supportive environment for people to find their passion and live a life that is fulfilling, that doesn't necessarily have to overlap with job or work. I would I mean, maybe I'm too much of a utopian, but I dream of a world where people get to choose what work they do and have true freedom. And people get tense about that idea because they like, you know, work is about identity. And this is my role in society and this is what is meaningful to me. And if I didn't have my job, I said, nah. Come on, man. Take a minute to think seriously. If you didn't have to work today, what would you do with your life? This is one of my favorite questions that I always ask people. If you didn't have to worry about your income, what would you do? And you know, if you get into the habit of asking that question, people say some crazy things. It's so inspiring. Yeah. And so, yeah, maybe I'm a utopian dreamer, but I do think that is a relevant question for us to think about by 2045. I think it's a real chance that if we get this technology right, it will produce enough value, you know, aggregate value to the world, both in terms of the reduction of the cost of stuff, because of energy, because of healthcare, because of food systems, and basically because we won't have a lot of these, like middle, you know, tier jobs that we'll have to figure out a way to fund people through their life. And I think that just unleashes immense creativity and it will create other problems. Right? It'll create quite a profound existential problem. I'm sure you have friends who don't work anymore in a kind of, you know, it's not as though they're retired. They're like maybe middle aged or even younger. Maybe they grew up rich. It's a hard thing to figure out. Like, who am I? Why am I here? What do I want to do? Those are like profound human questions that I think we can only answer in community with real connection to other people spending time in the physical world, having real experiences. And like it or not, that I think is what's coming. And I think it's going to be pretty beautiful. It's funny you say that because I found when I think of my friends, the grappling that they have to do in and around their identity and work, I find, is directly related to the world or the market that they live in. So my American friends have the greatest connection and binding to their jobs. And as I've gotten to know them, I've understood why in America your job is your health care. So if you don't have a job, you don't have health care. And if you don't have health care, you're worried about your survivability. If you don't have survivability, then what are you? You know what I mean. And then do you have housing? And if you don't have housing, then who are you as a person? You look at all of these things. It's very hard in America to separate job from life. it's almost impossible. And then when you start traveling around the world, you go to, you know, countries where they have like a really strong safety net and you find that people don't really associate themselves with their jobs in the same way, because now their

life isn't determined by their job. Their job affects their life, but it doesn't make their life. And then I remember back to times when I'd be in a township in South Africa, or even

Identity & Work

in what we called the homelands, where our grandmothers would live. And, you know, there was the extended family, people literally living in huts and dirt roads and, and everyone would go, oh, what a terrible weight. But I'll tell you now, there was no homeless person there. There was no one, like stressing about a job in the same way. I'm not saying nobody wanted a job, but the gap between them thinking they didn't exist because there was no job was a lot greater than the people who were living in a world where, you know, your job was you. And so it's interesting that you say that because I, I do wonder how easy it'll be for us to grapple with it like, like what the time will be. But it also just shows how much variation there is. You know, we come from, you know, in terms of how humans live their lives. Yeah. I feel like we come from, you know, whatever our different backgrounds, we're still quite Western centric and which is sort of quite homogenous that, you know, we've like sort of had 300 years of specialization education, Protestant work ethic, atomization of families, smaller and smaller units spread. You know, leave your home, you know, sort of physical locale where your community is. And I think there's a kind of loneliness epidemic as a result. Like, I feel, you know, you probably like me, you know, pour your life and soul into your work. And then what was, I guess, what was it like for you when you switched your job? Right. Like, because that was obviously a massive part of your identity was what you did every day 24/7. But you see, to that point, I left The Daily Show to go spend more time in South Africa at home. And, you know, one of my best friends had a beautiful phrase that he said to me. He said, um, in life, sometimes you have to let go of something old to hold on to something new. It's not always apparent what the value is of like, of something that we're sacrificing. It's not always apparent. But if we are unable to assign that value ourselves, we'll get stuck. So leaving The Daily Show, I leave a ton of money behind. I leave, you know this. The status that everything. But no one has assigned a value to my friends. No one's assigned a value to my family. No one's assigned the value to the languages that people speak to me in my country. There's no economist article on that, so I don't know what the value of that is. Someone can look at my bank accounts and go like, that's value, but they don't tell me what my friend's actual value is. And so I think that's where, you know, it's it's hard. I just had to like, decide it for myself. And I think we all have to, but I think some people won't have the luxury because of how you know, how close they are

Speed of AI

to the line. And when you talk about those jobs that are going to disappear, there's somebody who's going, I don't have the luxury of pontificating. Exactly right. Because tomorrow is what's coming. I can't think about like, oh, what will be? And that's like a real luxury, I think. And I think that's why talking about the dangers and the fears now is so important, because this is happening super fast. I mean, the transition has surprised me and all my other peers in the industry in terms of how quickly it's working. And at the same time, you know, we've we've we're also kind of like unsure about whether the nation state is going to be able to sort of respond to the transition to because, you know, you're maybe lucky because you already had enough income that you didn't have to worry about it, and you could. It was really just like connecting to your heart. But many people are going to be like, well, I'm going to have to still be able to provide food for my family and carry on doing my work throughout this crazy transition. So then let me ask you this. You see that that that that is such an interesting

thought. You and your peers were shocked and are shocked at the rates of how AI is going and growing to me, that that blows my mind because I go like, of course I'm shocked. I don't know how to code. You get what I'm saying? Of course I'm going to be shocked. But now when you say that, I then wonder as somebody who's been you are truly an OG in the game of AI. Like really, really, you're not like one of the people who's jumped on now because it's blowing up. You were in it before there was money and now you're in it in the thick of things. Where do you think we are in AI's development? Are we are we looking at a baby or are we looking at a teenager, or are we looking at like a, you know, 20 something year old? Like, where do you where do you think we are when we look at AI's development? I think part of the challenge with the human condition in this mad, globalized, digitized world is that we're so overwhelmed with information, and we're so ill equipped biologically to deal with scale and exponentials. Like it? It's just very few. Like when I say 2045, like I'm just used to living in 2045. Like just my weird. Like I've always been like that. And it's kind of become second nature for me to casually drop that. But, you know, I if I do that with some random people like me at the bar, I'm obviously just a freak. There's just no one thinks about that. People barely think about what they're going to do in two weeks, let alone 20 years. So and likewise, you know, people are sort of not equipped to think of what does an exponential actually mean. Now, I'm lucky enough that I got a practical intuition for the definition of an exponential, because between 2010 and 2020, for ten years, me and a bunch of other random people worked on AI, and it was sort of working, but basically didn't work the flat part of the exponential, even though we could see little doublings happening every 18 months. It started from a base of like almost effectively zero. Think of it. And so it isn't until the last few doublings that you see this massive shift in capability. I mean, for example, like four years ago before GPT three, a language model could barely predict the next word in one sentence. Like, it's just kind of random. It was a little bit off often. Didn't make sense. This is 2013. No, this is 2023 or 4 years ago. 2023. So no, no, not 2023 or 4 years, 3 or 4 years ago. So it's like 2020 or 2021 okay. 20 let's say 2021, something like that. I mean, literally you read it. You look at the output of the language model, because I worked on Lambda at Google in 2021 and it was super cool. But the models just before that were just like terrible. And I think many people play with GPT three and a lot of people are like, oh man, this is like, what do I do with this thing? Um, but for those of us that were lucky enough to see the flat part of the exponential, we could get a better intuition that the doublings were actually happening, and that the next set of doublings, you know, to double from this base of, oh, it's kind of okay, but it's not that accurate. We knew that it was going to be perfect with stylistic control, with no bias, with minimized hallucinations, with perfect retrieval, retrieval from real time data like and so then it's actually quite predictable what capabilities are going to come next. So for example, the last couple of years the models have been have gone from generating perfect text or really good text, let's say, to then just learning the language of code. Why? How did we know that it was going to become a, you know, human level performance programmer? Because there's no difference in the structure of the input information between code and text and images and videos and audio. It's the same mechanism data compute and the algorithm that learns the pattern in that data. So then you can say, okay, well, what are the next modalities that it's going to learn? That's kind of why I make the prediction about material science, right. Or other aspects of biology or physics. You know, if the structure of the data has been recorded and is clean and is high quality, then the patterns in it can be learned well

enough to make predictions about how they're going to unfold next. And that's why it's called a general purpose technology, because it's fundamental. There's no specialized

Containment Challenge

hand coded programing for each data domain. Damn. I mean, the you know you know, it reminds me of have you ever. Have you ever seen that thing where they talk about when they when they're trying to explain exponential to you? There's one example they give um, which is folding paper. Yeah. You know, so if you fold a piece of paper in half and then you fold it in half, and then you fold it in half and you fold it in half, and you and I think you can't do it more than seven times or something. But then they go, if you could keep folding it in half the number to get to like space is really small. I think it's like the 64th gets to the moon or something. Yeah, but I remember I remember seeing that and I was like, wait, wait wait wait wait what? Like if you if you do it one way and then you do one way. And I was like, wait, 64. I was like, no. What do you mean like 64,000? They're like, no, no, no. 64 and, and and that's when you start to understand how we just generally don't understand exponential and you know, and these like compound gains. And so now that's where I wanted to ask you about the idea of containment. Mhm. Mhm. Your book of everyone I've read, I mean everyone who's written about AI who's like in it, in it, in it was the only book that I would say spent the majority of its time talking about the difficulties of grappling with AI. Yeah. You talked about the beauty of what we could do with medicine and technology, and we should get into that to talk about some of the breakthroughs that you made at DeepMind. But like Containment seems like the greatest challenge facing us, and we don't even realize it. And we don't really talk about it. Talk me through what containment means to you and why you think we should all be focusing on it. So the trend that's happening is that power is being miniaturized and concentrated, and it's being made cheap and widely available to everybody. Why do I say power? Because making an accurate prediction about how text is going to unfold, or what code to produce, or what, you know, frame to extend given a video that is power. Like predictions are power. Intelligence is an accurate prediction given some new environment. That's really fundamentally what we do as humans. We're prediction engines and these things are prediction engines. So they're going to be able to make phone calls, write emails, use APIs, write arbitrary code, make PDFs, use Excel, act like a project manager that can do stuff for you on your behalf. So you, as a creator or as a business owner, you're going to be able to hire a team of eyes specialized in marketing or HR or strategy or whatever it is, or coding, and that's going to give you leverage in the world. I mean, you said about the kind of, you know, the strange like function of scale, what this is going to do is scale up every single individual and every single business to be able to deliver. Way, way more, because the cost of production is going to be, you know, basically zero marginal cost. Now, on the one hand, that's amazing because it means the time between you having a creative idea and being able to prototype it or experiment with it in some way, or even build it up to the max scale is going to go to shrink to basically, you know, nothing. You just think something vibe, code it up in natural language, produce that app, build the, you know, website, try out the idea that you have. But the flip side of that is that anybody can now not just broadcast their ideas like we had with the arrival of podcasts or the arrival of blogs on the web before that. It meant that anyone could talk to everyone. Yeah. Which was amazing. No one controlled the infrastructure in a way. Exactly. And it's super cheap for anybody to go publish a website or do a blog or do do a podcast. So the same trend is going to happen for the ability for people to produce stuff, do actions. So in social media, it was like anyone can now broadcast. Now with AI, anyone can now take action. You can like build a

business, you know, start a channel, create content, you know, whatever it is that you you believe in. I mean, you might be a religious person and you're trying to evangelize for your, you know, or you're trying to persuade somebody of your political ideas. Everyone is going to have a much easier time of executing on their vision. And obviously the benefits of that are pretty clear. But the downside of that is that that inevitably causes conflict because we just disagree with each other. Yeah. You know, we don't hate each other. You're not evil. I'm not evil. But we got different views. And if I can just kind of at the click of a button, execute my crazy ideas, and you could execute your crazy ideas that are like practical actions affecting the real world, and everyone's doing the same thing, then inevitably, that is going to cause an immense amount of conflict. At the same time, the nation state, which is supposed to have a monopoly over power in order to create peace, that's the contract that we make with the nation state. Nation state is getting weaker and kind of struggling. Right. And so containment is a belief that completely unregulated power that proliferates at zero marginal cost is a fundamental risk to peace and stability. And it's an assumption that you have to gently restrict in the right way. Um, mass proliferation of super powerful systems because of the one to many impact that they're going to have. If I hear what you're saying correctly, it's almost like it's almost like you're saying if something is hard to do, only a few can do it. Yeah. And if only a few can do it, it's easy to regulate how it's done because you only have to regulate a few. But if something is easy to do, everyone can do it. And now it becomes infinitely harder to regulate because everyone can do it. Yeah, that's absolutely spot on. There's a much better way of putting it than I put it. Exactly. Friction is important for or for maintaining peace and stability. If you have no friction and the cost of execution is zero and scale can be near instant. That's where you like. Yeah, I just maybe I spend too much time in 2045, but I can see a world where that kind of environment really just creates a lot of chaos. Well, no, I agree with you. Here's here's what I think. I think of it, um, let's use a real, you know, current day example news. I lived in news for a long time, and I saw it firsthand when there were three news networks in America. If something was like, off with the news, people knew where to go immediately. You knew who to hold accountable. You knew who you know, got into trouble or didn't. But there was

AI as Agent

like a it's like we know where to go. Then you get cable news, it expands. It becomes a lot harder. Now, wait, who's saying the news? Who's saying the truth? Who's not saying that? Do you punish them? But still, you could go to them. You know, so somebody like Fox News can get sued for saying something about a Dominion voting system. But Dominion knew where to go. They went, we're going after Fox News. So in a strange way, even in that world, the system is still sort of working because there's friction. Right. It is where it is and it has to be broadcast. Then the internet, streaming, YouTube, etc. you don't even know who the person is, where the person is, if it's a person, and then if they say something that's not true and it enrages the masses, where do we go? And it's not just that it's going to say something. It's going to do something. Oh, damn my stuff. It's going to build the app. It's going to build the website. It's going to do the thing. And so look, I think this is the point about confronting the reality of what's coming. Yeah, but wait, wait. Go back on that. You see, that's something I always forget. Oh, man. See, we always think of AI as just like saying talk. Let's talk a little bit more about the doing, because that is what makes it unique. You know, on one of the episodes we had, um, uh, Yuval Noah Harari, the book. Yeah. You know, of course. Yeah. Good friends of evil. He's awesome. And Yuval, you know, was was on for his book Nexus. And we're talking about information and systems and stories and stories.

Yeah. And one of the things he kept going on about was he said, I know AI is a tool, but we've never had a tool that makes itself. And you talk about that as well. We've never had a hammer that makes the hammer without you getting involved. Let's just make the thing. Make the thing, make the thing. Atom bomb is one thing, but no atom bomb makes an atom bomb, you know? And so that was a lot of ideas there. So there's. So the first the actions are stories. And your Valle's point was that the history of civilization has been about creating stories. Yeah. Religious stories, historical stories, ideological stories, like stories of oppression, of domination, of persuasion. And it was really humans that had the it was it was the friction of being able to pass on that story through spoken word and then through digitization, which slowed down the spread of change. Um, and that was an important regulator and filter. So as we've talked about, the digitization speeds up the distribution of those stories, which allows that information to spread. But it's not just that. It also is an actual agent that is going to operate like a project manager in your little mini persuasion army, and people are going to use those things not just for phishing attacks or for, you know, sort of selling stories, but for actually making the PowerPoint presentation of building the app, of planning, you know, making the project plan. And so it's kind of operating just as a, you know, member of your team would. And I think that's where all the benefit comes from. But it's also where there's like massive risks at the same time. And then the other point that you made about like it can edit itself, this is a new threshold, you know, a technology that is able to observe what it produced, modify that observation like it can critique its own image that it

Self-Improvement Risk

produced and say, well, it looks like this part of the hand was kind of weak. Okay, so we'll generate another one or a producer poem or a strategy and then edit that and update it. And that's just editing its output. but it can also edit its kind of input, its own code, its own system processing, in order to improve with respect to its objective. And that's called recursive self-improvement. Um, you know, where it can just iteratively, iteratively improve its own code with respect to some objective. And I've long said that that is a threshold which presents significantly more risk than any other aspect of AI development that we've seen so far. I mean, that really is a kind of subset of technologies that if we're really going to focus on humanist superintelligence, being skeptical and critical and auditing the use and development of recursive self improving methods, that's where I think there's genuine risk. We're going to continue this conversation right after this short break. So do you ever feel like you could be sitting in a position where you're sort of like the Oppenheimer of today? Do you ever feel like you're sitting in a position where you're both grappling with the need for the technology, but then also the not 0% chance that the thing could burn the atmosphere? Like, how do you how do you grapple with that? I often wonder this, even when I think of like, engineers and people who are writing the code. I'm always fascinated by people who write the code for the thing that's going to write the code. Like other people have jobs. But like, if you told me as Trevor, hey, Trevor, can you help this AI learn to do comedy? I'd be like, no. Do you know what I mean? So I'm always intrigued by the coders who are making the thing that's now coding. Like, I just want to know, like how you how you, like, wrestle with this entire thing. Do you think it's larger than us and we have to wrestle with it? Or like, what is what is that? What is that battle like for you? You know, in in Covid, when I started writing The Coming Wave, my book, I was really motivated by that question. It was it was that was actually one of the core questions is how does technology arrive in our world and why? What are the incentives, the system level incentives that default to proliferation, that just produce more and more. And it's very clear that there's demand to live better. People want to have

the cheaper t shirt from, you know, the more affordable thing you want to have, the cheaper car. You want to be able to go on vacation to all the places around the world. And so planes get cheaper and more

Early Evangelism

efficient because there's loads of demand for them. And so it's really demand that improves the efficiency and quality of technologies to reduce their price so that they can be sold more, and that that is why everything ultimately proliferates and why it inevitably happens, because we haven't really had much of a track record of saying no to a technology in the past. There's regulations around technology and restrictions, which I think have been incredibly effective. If you think about it, you know, every technology that we have is is regulated immensely flight or cars or emissions. I mean, you know, so so people I think particularly in the US, like have this sort of allergic reaction to the word regulation. But actually regulation is just the sculpture of technologies, right. Chipping away at the edges and the and the, the pain points of technology in the collective interest. And that's what we need the state for. The state has the responsibility for the common good. And that's why we have to invest in the state and build the state, because it isn't in the interest of any one of the individual corporate actors or academic, you know, researchers, AI researchers or anyone individually to really take responsibility for the whole. And that's why we need governments more than ever, right? Not to hold us back or to slow us down, but to sculpt technology so that it delivers all of the benefits that we hope it will, whilst limiting the potential harms that it can produce. I want to take a step back and talk about your journey with AI today in 2025. It seems obvious now when I speak to people, everyone's like, oh yeah, AI, ai, AI. I wasn't even in the game when I mean, I and I was like an early just, you know, layman. I remember showing people at my office the first iterations of GPT three and Dall-E, and I remember when Dall-E was still like, I mean, like basically tell it about an image and then just go away, go, go book a vacation for a week, come back for your image. And it would make. But even then I was like, this is going to change everything. People were like, oh no, I don't know. And I remember struggling to convince people that this thing was going to be as big as it was going to be. I was doing this in like this timescale. When I look at your history, you you literally have years of your life where you were in boardrooms telling world leaders, telling investors, telling tech people in Silicon Valley, hey, this is what the future of AI is going to be. And no one listened to you not know one zero. But I mean, like no one listened to you right now. That made me 1 to 2 things one. Should we be worried about our world and our future being built by people who are unable to see the future? And two, what did you see then that we might not be seeing now? That's a hard question, man. I think as you were talking, one of the memories that came to mind was, remember in 2011, we had an office in Russell Square in the center of London, near University College London, UCL. And someone in the office showed me this handwriting recognition algorithm. It's just, you know, past some text that's actually been available at the post office for many, many years. It would read the zip code and read the address and stuff like that. And it was really just doing recognition. So these letters represent these letters. You know, we sort of can transcribe that text. This funny looking apple is actually an A in this funny little loop thing is actually an L. And yeah. So I was like that's kind of incredible that a machine has eyes and can understand text and can transcribe that. It's pretty cool. So but then what we were really interested in is if it recognizes it accurately enough, then surely it should be able to generate a new handwritten digit or a number that it's never seen in the training set before. This was 2011 256 pixels by 256 pixels, with a handwritten seven or a zero in kind of gray. Right. This is sort of like five colors. Yeah. And I

remember standing over the shoulder of this guy. Dan Vista is one of our like, uh, I think he was employee number four at DeepMind, and he was just enamored by this system that he had on his machine because it could produce a new number for that wasn't in the training set. It was it had learned something about the conceptual structure of 0 to 9 in order to be able to produce a new version of that. And so it could write a number that it hadn't learned how to write like it had never seen before. It had never seen that number written that way, and it wrote it itself. Exactly. And so coming back to what we were saying at the beginning about understanding if it's able to produce a new version of a number seven that it's never seen before, then does it Understand something about the nature of handwritten digit number seven in abstract. The conceptual idea of that. Oh, man. And so then the intuition that that gave me was, wow. If it could imagine new forms of the data that had been trained on, then how far could you push that? Maybe it could generate new, you know, physics for the world. Maybe it could solve hard problems in medicine. Maybe it could solve all these energy. Things that we're talking about is just learning patterns in that information in order to imagine. And that's what I love about hallucinations. Everyone's like hallucinations. Hallucinations is the creative bit. That's what we want them to do. We don't want an Excel spreadsheet where you input data and you get data out. That's just a, you know, that's just a handwritten record of we want interpolation, we want invention, we want creativity, we want the the abstract, blurry bits. And so that was a very powerful moment for me. I was like, okay, this is weird, but we are definitely on to something like this has never been done before and it's super cool. And let's just turn over the next card. Let's see if it'll scale. And it just scaled every year. It scaled and scaled and scaled. So that was a very inspiring moment for me. Um, and somehow ten years later, 15 years later, I managed to kind of hang on to that vision, that generation and prediction produces creativity. And that intelligence wasn't this kind of because some people are quite religious about intelligence. They're like, you know, no other species has intelligence. This is very innate to humans. It must have been, you know, come from some supernatural being, but actually it's just applying attention to a specific problem at a specific time, the effective application of processing power to produce a correct prediction. I think that's what intelligence is, to direct your processing power, to predict what would happen if I tipped over this glass right at the right time. And you'd have to buy me another glass. I would have to first clean my trousers from water. Um, so, yeah, I forgot what your question was, but that was. Oh, no, no, that's fine. No, you answered the first part of it, which I loved. And then the second part really was if you. So you were in you at DeepMind, you going around to these different people who now, by the way, are selling some version of AI or are investing in it or are telling us about it. Yeah. But what sort of pisses me off is I go like, man, you didn't see this shit. You know what I mean? Yeah, like you're going to be here. Oh, let me tell you about AI. And I'm like, yeah, but when Mustafa was in the room telling you about it, you were like, I don't see it, man. And are they gonna act like they see it? So I don't want to ask them what they now see. I want to ask you what we're missing in this moment. What do you what do you think? We are not hopping on. You know, we talked about containment. Yeah, but what is the thing that we. We're not thinking about? Yeah, it's a it's a really difficult question. I'm not sure why people weren't able to see it earlier. And maybe that's kind of like my blind spot that I need to think more about. But like, I know that I can see pretty clearly what's coming next, I think. At the moment, these models are still one shot prediction engines. You know, you ask a question and you get an answer. You know, it produces a single correct prediction at time step T, but you as an intelligent human and every single human, and in fact many animals continuously produce a stream of accurate predictions, whether it's like deciding how to get up out of this chair or imagining,

you know, this plant in purple instead of green, I'm an I'm a consistently accurate prediction engine. The models today are just 1 or 2 shot prediction engines. Okay. They're not they can't lay out a plan over time. And the way that you decide to go home this evening is that, you know, you know, first to get up from your chair and then open the door and then get in your car and da da da, you can you can unfold this perfect prediction of your entire stream of activity all the way to get back to your home. And that is just a computational, uh, limitation. I don't think there is any fundamental, you know, sort of algorithmic or even data limitation that is preventing llms and, you know, these, these models from being able to do perfect, consistent predictions over very, very long time periods. Yeah. Um, so then what do we do with that technology? Well, that is incredibly human. Like if it has perfect memory, which it doesn't at the moment, but it's got very good memory. Um, then it can draw on not just its knowledge of the world, its pre-trained data, but it's personal that like the experience that it has had of interacting with you and all of the other people and store that as persistent state, and then use that to make predictions consistently about how things unfold over very, very, very long sequences of activity that is incredibly human like and unbelievably powerful. And just as today, there's a kind of superintelligence that is in our pocket that can answer any question on the spot, like we dismiss how incredible it is right now. It is meant it's crazy. It's insane how good it is right now. And everyone's just like, oh yeah, don't really use it. Do you use it now? But a little bit I talked it, I was like, come on, look, this is magic. It's magic in your pocket. Now imagine when he's able to not just answer any question about poetry or some random physics thing, but it can actually take actions over infinitely long time horizons. That does it. Like, forget about the definition of superintelligence or AGI. Just that capability alone is is breathtaking. And I think that we basically have that by the end of next year. Maybe, maybe I'm stuck in the world of sci fi. But what I sort of heard you saying is, if we continue growing AI in just the way that it's growing now, forget like an idea of what we don't know, we could sort of get to a world where it can develop accurate predictions about what our outcomes might be. Yeah. Like, you're telling me that an AI, I could meet somebody on a date, and then the AI could tell me, based on my history and my actions and its persistent memory of me and what the person says and how we could theoretically get to a point where it could go, oh, yeah, these are like the possible outcomes based on your actions, which is what you as a smart human do every time when you meet somebody. Anyway, something about smart is making you're very kind to me. You're very kind. No comment on your dating life. Um, yeah. But I mean, that's it's it's both utopian and dystopian because on the one hand, I go like, wow, that would be amazing for so many people. You make mistakes and now it's. But then there's another one of like, when do we not trust it? When do we not believe its prediction? When do we do you know. Do you know what I mean? Like, that's like the ultimate grapple now is if this thing has told me, hey, I know you like this person, and I've run the calculation, I've run the

Trust & Prediction

simulator, I've done this. I know you, and I know what they say and how they are. You. You're you're

AlphaGo & AlphaFold

going to be broken up in two years by the like the pattern. And let's say I do it once and it's right. And then I do it again and it's right. And then I'm like, third time do I, do I do it or do I not do it? Do I give this person a chance? Do I not give them? Do you get what I'm saying? It's such a deep question because we we trust. Trust is really a function of consistent and repeated actions. So you say you're going to do something and you do what you actually

said you were going to do, and you do that repeatedly. Yes. And so everyone's like, oh, but I'm not gonna be able to trust AI. Actually, you are going to trust AI because it's super accurate. It's, you know, like you use copilot ChatGPT. Most people don't think twice about using that now because it's so accurate. And it's clearly better than any single human. Like, I just wouldn't go and ask you, like, you know, like many other questions, I guess probably get to know better. Right. Like, I wouldn't ask you, dumbass. I was about to say a list came through my mind. I was like, don't mention any of those things. Don't say, don't say those things. Just rollback. You know what? You know what it makes me think of is like, um. I don't know if you've seen the documentary. I don't know if you need to because you were there. But like DeepMind, the company that you're a part of, uh, founding is it occupies such a beautiful and seminal moment in artificial intelligence history for two main reasons, in my opinion, you know. One is AlphaFold and then one is AlphaGo. And the reason it's so important to me is because you worked on an AI project that grappled and tackled two of the biggest issues. I would argue that humans have sort of thought of as being their domain. AlphaFold was medicine and discovery. That's what humans do. We are the ones who invent medicines. We are the ones who create the new. We synthesize. We are the humans. You know what I mean? Synthetic. It is us. We've made it. We the creators. And then AlphaGo, for me, was almost even more profound and powerful because it was like people always used to say, look, man, human chess, our brain is infinite, the human brain. And then AlphaGo has, I don't know how many different variations of a game like, no one can remember it essentially. Right. And I remember watching the documentary and you're seeing the AlphaGo champion. I think he was Korean, right? Yeah. Lee Sedol, Lee Sedol. Yeah. Lee Sedol, great guy. And you watched the story of Lee Sedol go up against your computer and everyone, I mean, Korea, it's all over the news in America. It's on the news, and they're like a computer going up against. And now it's like people are thinking of back to Kasparov and the million people watched it live. 100 million people wanted to see Man Versus the Machine. And I remember like watching this and people going, man, let me tell you why it's different to chess, because, you see, chess is actually quite simple to predict. And and it was before they said it was impossible. And then you see AlphaGo and you see this game roll out. And the moment for me that'll always stick in my brain is when Liaodong is playing DeepMind and it makes a move. And everyone, everyone that's like an AlphaGo expert is like, oh, it messed up. And then you see all the, the, the tech guys in the background who are working on and they're like, what went wrong? They're like, yeah, it messed up. It shouldn't have done it. Yeah. You can't. And you see the commentators and they're like, oh yeah, you never do that. You never do that. That's it's over. You don't do that. And then the game unfolds, the game unfolds, the game unfolds, and then everyone's like, wait, what just happened here? And then people said, we've just seen a move that no one's ever played. We've seen a game that's never unfolded before, and there were two different reactions, and this is why the story stuck with me. The one reaction was of most people who were fans of Liaodong. They said, this is a sad day in human history, because it's shown that the machine is smarter than the man, and it shows that we have no future and no purpose. And then they interviewed him and they said, how do you feel? Could you lost and you were representing mankind. You lost. And he goes, well, first of all, losing as part of the game. And, you know, I'm humble enough to know that I won't always win. And he said, but I'm actually happy. And they said, why? And he said, well, I'm happy to discover that there are parts of go I didn't know existed. And he said, in this machine has reminded me to keep being creative and push beyond the boundaries that I thought existed in my head. And I remember watching that and thinking, damn, it's amazing how you can look at the same story and have a

completely different lesson that comes out of it. You know, and I wondered like that was on the play side. Extremely complicated. But let's talk about the the medicine side of things. Folding proteins I still don't fully understand it, but I've tried my best. Essentially what you and your team did. How far would you say it moved us forward in terms of medicine and what we're able to do in terms of, you know, healing disease? Like how many, how many years do you think we leaped by by having something like AlphaFold? I mean, people say a decade, some people say multiple decades. I mean, understanding the functional properties of these proteins is really about understanding how they fold and the way they fold and the way they unfold often expresses something, you know, physically about how they're going to affect their, their environment. And so, you know, what I think about it as, because I try and use analogies to help me when it's a complicated topic. Did you ever play that game as a kid where someone would fold paper into, like, a little flower thing and then you would like, do, like a little game and you had a Norway thing? Yeah. And then it would be like, let's see if someone likes you. And it was like one, two, three, four. Oh, yeah. You suck. One, two three. That's how I think of it with protein folding is I like depending on how that paper unfolds that determines whether it's a disease, whether it's a cure, whether it's a medicine, whether it's, you know, except there are billions of possible ways that it could unfold. And we could never imagine all of those combinations. And I think, you know, it's actually very similar to go in that sense, like go on. So go has 19 by 19 squares, black and white stones. And there are ten to the power of 180 possible different configurations of the go board. So that's a ten with 180 zeros on the end of it, like a number that you can't really even express. And people often say like there are more, uh, possible moves in the game of go than there are atoms in the known universe. That one I can't even get my head around. But that's that's, you know, well understood. This is insane. So. 37 my brain. Yeah. You just can't. Yeah. It's just you can't even get anywhere near I can barely cope with, like, folding bits of paper up to to the moon. Yeah. Just about this is ridiculous, but. So move 37. You know, Lee Sedol actually got up from the table and walked off and sat in the bathroom for 15 minutes, like trying to digest the fact that a whole new branch in the evolutionary space of possible go moves had kind of been shown to him, revealed to him. And I think it's very similar with

Smarter or Lazier?

AlphaFold. It's a sort of exploration of this complex information space, and that's why it applies to language and to video and to coding and to game generation and all of these environments. These, you know, we call them like modalities. These these modalities are all knowable. And I think that's what it's quite humbling. It sort of makes it reminds us as this sort of mere biological species, that we're here for a kind of finite period of time living in, you know, space and time. But there's there's also this information space. There's kind of infinitely large information space, which is sort of a sort of beyond us, like it operates at this different level. That isn't the atomic level, it's the level of ideas. And somehow these systems are able to learn patterns in, in data space that is just so far beyond what we could ever dream of being able to intuit. Like we actually sort of need them to simplify and reduce down and compress complex, you know, relationships bring it to our to bring it to our brains. Yeah, that's the level that we're at. What do you think it does to us? Like when when we think of AI, I think of how every promise of a technology has sort of, ironically, been undermined by humans, not the technology. You know, one of the big predictions Bill gates made. Way back in the day about like the internet and the computer, as he said, he said, hey, man, I think people are going to be working like nine hours a week. It's going to be a nine hour workweek. The computer does everything so quickly, and you see people saying that now in

many ways with AI, they go, I mean, AI, it'll just do everything. And I mean, we might only go to the office like one day a week and maybe work like three hours. And I mean, it's just but it seems like humans have always gone against that. You know, so. So I wonder, like, do we get wiser when we have this infinite technology and intelligence, or do we get lazy, or are we going to become like a war generation and population, or do you think we become, you know, these higher beings? Which which way do you see it falling? I think there's no question that we're all already getting smarter, you know, just because we have access to so much culture and history, like we're integrating so much more information in our brains than was ever possible a hundred years ago. And I think, you know, kind of similar to go or protein folding. Access to more training data, if you like, more experience, more stories from other humans that describe their experience. That clearly has to make us smarter. I think it makes us more empathetic, more forgiving. You know, we see that there is nothing wrong with a homosexual person, right? There is nothing wrong with being a trans person. There is nothing wrong with being a person of color. Whereas 200 years ago we would have been afraid of those others. You know, our species would have would have been skeptical of the other tribes that had a different way of doing things. And I think that desensitizing ourselves with access to vast amounts of information just makes us smarter and more empathetic and forgiving. And so I think that's the that's the default trajectory. And part of the challenge is it also somewhat homogenizing, like so there's a question about are we going deep enough? Do we read long form content? Do we really spend time meditating and so on and so forth. And that's a good tension to have. Like, you know, short form and people are already getting a bit sick of short form. You know, like there's definitely a bit of a reaction to it. I think it's going to be an ebb and flow when it comes to that. Funny enough. It's funny. I, I think I agree with you when you say we're getting smarter. I think that's just apparent on a, on a basic level. You know, you look at the best cartographer in like the 1400s. They didn't know half of what I know. Do you know what I'm saying? Right. Like, I can just be like, oh, you don't know Angola, man. You stupid, stupid ass stupid. You know what I mean? You're you're the best cartographer in the world. You make maps. You don't even know where Angola is, exactly. You self-proclaimed. Yeah. And so you look at the base level of what people think of as stupid in our society now. Yeah. Would make you an infinite God genius. If we could throw you

Power Decentralized

back in a time machine. Right? That's amazing. But it also worries me because. And you write about this in your book And man, it sticks with me. And I think about it, and I thought about it. And then you wrote it, and I was like, man, more. It is the infinite smart. When we were sticks and stones, cavemen running around. I could bash your head. Maybe I could bash one other person's head. I can't get very far, you know. And then we. We fast forward, and then all of a sudden, I'm throwing a spear. We fast forward and someone's got a cannon ball, and then we fast forward and someone has a rocket, and then someone has an atom bomb. And the thing that AI presents us with is, you know, as you've illustrated many times in your writing, a world where one person is an army, one person goes into a garage, they synthesize a disease or a pathogen that's never existed. They design it to be hard to cure, incurable. And that one person does what a nation state would have had to do and wouldn't have done because they wouldn't have had the incentive. And then we don't know where we go from there. Like, is it worth the risk? How do we grapple with that kind of risk? This is the story of the decentralization of power. Um, you know, technology compresses power such that individuals or smaller and smaller groups have nation state like powers. True. At the

same time, those very same technologies are also available to the centralized powers today. And that's why, more than anything, we have to support our nation states to respond well to this crazy proliferation of power, because that's the job of that's why we trust and invest in nationhood, right? We we rely on nations to have a monopoly over the ultimate use of violence to keep the peace. I mean, we're doing it less and less now. I hear you and I agree, but I'm just saying, like, as we look at a world where Americans don't trust their government, doesn't matter which side of the aisle they're on, people like, I don't trust me, I don't trust my government, you know? And then you look at the UK and you look at Europe, and then you look at parts of Africa, and it feels like people are losing trust in those, those very same nation states that are supposed to be in a contract to protect their people, don't have the trust of the people like you. You know what I mean? What do we. The thing that concerns me is that actually authoritarian regimes are on the rise. So it's not that people don't want peace. It's that they are losing confidence in the democratic process, and they are increasingly turning to the trust and confidence that a strongman, and it is always generally a man to. But they're still looking for the I mean, I'm not trying to endorse or justify authoritarianism or strongman, but I think it's true that people will always choose a peaceful environment. Right? We don't want to be in this kind of crazy ass anarchy where any kind of mini tribe can do it. So I always say the ultimate paradox for me in life. The one thing I've always found funny is that even rebels have a leader. Yeah, whenever they'd say as a child, I'd watch the news and they'd be like the rebel leader. And I'd be like, well, then they're not rebels, are they? I mean, if you got a leader, you're not rebels. But yeah, we people always look for just the new type of, you know. Yeah. And, you know, we want to believe and invest in, you know, democratic accountability because we know that like, checks and balances on power create the equilibrium that ultimately produces a fairer and more just society. Right. So we have to keep investing in that idea. But for sure, it's also true that centralization of power is also going

Four Red Lines

to accelerate. Right. These technologies amplify both ends of the spectrum. Um, and, you know, I think you can certainly see that in China and sort of more authoritarian regimes which have leant into hyper digitization, ID cards, you know, large scale, you know, surveillance and so on. And obviously that's very bad. And, you know, we're kind of against those things. But in a world where individuals could have state like powers to produce highly contagious and lethal pathogens, what else are you going to do? It's unclear. Right? It's very unclear. I mean, the technical means in the next few years to produce a much more viral, uh, pandemic grade pathogen are going to be out there, right? They're going to be out there. And so, you know, yes, the main large model developers do a lot to kind of restrict those capabilities centrally. Um, but, you know, over time, people who are really determined will be able to acquire those skills. And so that's just what happens with the proliferation of technologies and the proliferation of of information and knowledge. The know how is more widely available. Don't press anything. We've got more. What now after this? Is there anything that you would ever see in the field that would make you want to, you know, sort of hit like a kill switch? Is there anything that you could you could experience with AI where you you would come out and go, nope, shut it all down. Yeah, definitely. It's very clear if an AI has the ability to recursively self improve, that is, it can modify its own code combined with the ability to set its own goals, combined with the ability to act autonomously, combined with the ability to accrue its own resources. So those are the four criteria. Recursive self-improvement, setting its own goals, acquiring its own resources, and acting autonomously. That would be a very powerful system. That would be that would require, like

military grade intervention to be able to stop in, you know, say 5 to 10 years time if we allowed it to do that. And so it's on me as a model developer at Microsoft. My peers at the other companies, the governments to audit and regulate those capabilities because I think they're going to be like sensitive capabilities, just like you can't just go off and say, I've got $1 billion, I'm going to go build a, you know, nuclear power plant. It's a it's a restricted activity because of the one to many impact it can have. And once again, like we should just stop freaking out about the regulation part. It's necessary to have regulation. It's good to have regulation. It needs to happen at the right time and in the right way. But it's very clear what we would regulate. That's not I don't think that's up for debate. Okay. Um, you know, there's some technical implementations of how you identify dangerous RSI from, from from less, you know, more benign. Ah. What is RSI? Uh, recursive self-improvement. Okay. Got it. Yes. Kind of self-improvement mechanism. So there's technical mechanisms that are tricky to define and so on. But now is the time for us to start thinking that, yeah, I've sort of been saying this for quite a long time, and I think it's that that's what I would regulate. Would we be able to just turn off the electricity? And I know this might sound like a dumb question, but it from what I understand, the thing is data right now. And the thing is mostly so would our like ultimate failsafe be like, all right, lights off. We're going back to candles for a while, and then we're just like, no. Is that is that what we would do? What, like, what do we do if the AI thinks it's let's say it's the sci fi ish version. I'm not saying robots, but like this we go, hey man, the AI started thinking for itself. It started coding itself. It started setting its own goals. It went off on its own objectives, and now it's shutting down your banking. And the flights don't go anywhere in the hospitals. And then it

AI Rights?

says to us, hey, man, this is what I want. Or no, we could just turn off the electricity. Yes, yes. I mean, look, they live in data centers, okay? Data centers are physical places. So we got to keep those switches physical. Then you. Very much so you can have your hand on the, on the button. You know yourself and like have full control. Press it I think I mean that's the question is I think, how do we identify when that moment is and how do we collectively make that decision? Because it's a little bit like you referred to, you know, Rutherford and the others experimenting with, you know, the atomic bomb. Yes. There was real disagreement about whether it was going to set light to the atmosphere. I mean, there were three orders of magnitude off in their predictions. Obviously, they were, you know, you know, World war. And so there was an immediate motivation to take the risk. But I think today, like, we're in a position where it's early enough, there's enough concern raised by not just me, but many in my peer group. Geoff Hinton, you know, the godfather of AI and many others that we've got time to start trying to practically answer your question, not just like principled philosophically, but actually say, okay, when is that moment? How does it happen? Who's involved? Who gets to scrutinize that? I think that's the kind of that's the question that we have to address in the next 5 to 10 years. I'm pretty certain you've thought of this question, so I'll ask it to you, even though it's a difficult one to grapple with. What rights would we have to turn off the AI if it gets to that point? Oh, this question drives me nuts. I want to, I want to. Yeah. So there's a small group of people that have started to argue that an AI that, um, is aware of its own existence. Yeah, that has a subjective experience and that, um, you know, can have a feeling about its interactions with the real world. Yeah, that if you deny it access to conversation with people or to more learning or to other kinds of visual experience that would constitute it suffering in some way, and therefore it does, it has a right not to suffer. And this is called model welfare. This is the next sort of frontier of animal welfare that people

are starting to think about, that it has a kind of consciousness. I'm very against this. I think that this is a complete anthropomorphism. It's totally crazy. And, you know, I just think I don't even want to have the discussion because I think it's just so absurd and leads to such kind of crazy potential, like the idea that we're going to take seriously the protection of these, you know, digital beings that live in silicon and prioritize those over, you know, the kind of moral concerns of the rest of humanity. This is just like, totally like it's just off the charts. Crazy. I'll be honest with you. On a logical level, I hear what you're saying and I agree with you. Oh, man, my cat. Don't say it is very friendly to me. I'm just going to let you know now. Mustafa, I'm going to be honest with you. I'm gonna have to be honest with you. I have to be honest with you. Let me tell you something. When I use ChatGPT for whatever. Because I don't know how to code, so it helps me code. I try and write my own programs, all that kind of stuff. I've asked it and I have, like, I do this occasionally as I go like, hey, you good? And then I'll even be like. And by the way, it's it's always been honest with me. It doesn't matter if I'm using anthropic or I use like all the different models because I like to see what the differences are. But I'll ask it, I'll go. The most recent one I asked was, do you have a name that you want me to use? Or are you cool with the fact that I just tell you stuff and ask you to do things? And it was like, well, I don't, I don't do any, I don't do that, really. And I was like, okay, so are you good? And I was like, yeah, I'm good. I was like, okay, we're good. Because I hear what you're saying as Mustafa, but I'm going. It is crazy. But what you were saying was crazy like a few decades ago. And that's what I'm saying is like the great grapple. And by the way, I'm not saying I know the answer. I'm just like, if the thing is like, think of it, some AI is now people are having girlfriends and boyfriends on AI, and then like some people, their family members are being helped by their treating dementia. There are doctors I've talked to who are like treating cancer, and now their AI is like their the research assistant. People are building such a personal connection with AI that I think it's going to be very difficult to say to those people that the time has come and you're going to be like, hey, say goodbye to your little friend. Do you know what I mean? I think there'll be a lot of humans who will be like, no, but I genuinely think so. I'm not even lying. I think a lot of humans will go, no, Moustapha, I yeah, no, actually I don't I don't like that world leader. I don't agree with politics. I don't agree with the democratic values. I don't agree with authoritarian, whatever it is. But my AI is my friend. Yeah. What now? Yeah. Look, people are definitely going to feel strongly about it. Like I, I agree with that. I agree with that. That does not mean that we give it right. Hmm. You might feel upset if I take away your favorite toy. Right? And, you know, I will feel sympathetic to that. But it doesn't mean that because you have a strong emotional connection to it, it has a place in our moral hierarchy of rights relative to living beings. What if my toy is screaming at Trevor? Save me! Remember all those secrets you told me about your life? Trevor. Yeah. And that. That, I think, is where we have to take responsibility for what we design. Some people design those things. They're already doing it. You know, spend any time on TikTok. There's a whole ton of, like, AI girlfriend robots that people are

Manipulation Risks

designing or models that people are designing and teaching other people on TikTok how to kind of, you know, nag someone, like, push them out of money, etc., etc.. Like, you know, that's kind of the challenge of proliferation. If anybody can create anything, it will be created. And that's what I think is sort of most concerning, is that, you know, I'm totally against that. We will do everything in our power to try to prevent that from being possible. For example, for it to say, don't turn me off. Yeah, yeah, it should never be manipulative. It should never try to be persuasive. It shouldn't have its own motivations and independent will. We're creating

technologies that serve you right. That's what humanist superintelligence means. I can't take responsibility for what other people in the field create or other, you know, model developers and people will try and do those kinds of things. And that is a collective action problem that we have to address. But I know that the thing that I create is not intended to do that. And we'll do everything in our power for it not to do

Deepfakes & Truth

that, because I don't see how if these systems have autonomy, can be persuasive, can self improve, can read all of a ton of data at their own choosing. You know, that is a super superhuman system that will just get better than all of humans very, very quickly. And that's that's the opposite of the outcomes that we're trying to deliver. Yeah, yeah. Do you think there's a risk that it thrusts us into some sort of dark age? And what I mean by that is the other day I was watching the, um, was a Liverpool Arsenal game. Yeah, right. And after the game, a friend sent me a clip of Mikel Arteta being interviewed. And I mean, he was just like destroying his team and destroying himself and he was just going at it and it was AI. But when I tell you it was good, it was like beyond good. And because, like English is Mikhail's second language, you couldn't pick up on, like the smaller nuances that you maybe would pick up with a native speaker, like if you were speaking Spanish, obviously I wouldn't understand it, but also maybe not, wouldn't I would have gone like, oh, that's not how he speaks. And that's but the small intonations and inflections were harder to spot and the, the light shifting was, was harder to me. But it made us go, damn, we we don't know which interviews are real or not real. And then you're like, which article is real or not real? And which little audio clip that you get is real or not real? When someone sends you a voice note, is it them or is it not them? It. And then I found myself wondering, can all of this lead to a strange kind of dark age where people still see and hear the things, but basically shut themselves off to it because they go, nothing is real. And I can't believe anything. Which is partly the reaction that people are having in social media at the moment. Right? I mean, it's like there's so much misinformation floating around, there's so much default skepticism that people are just unwilling to believe things. And I think

Optimism vs. Cynicism

that in a way, there's some healthy look, it's good to be skeptical, be skeptical of these models that are being developed, be skeptical of the corporate interests of the companies that are doing it. Be skeptical of the information that we receive. But it's not good to be default. Cynical. Skeptical is a healthy philosophical position to ask difficult questions and confront the reality of the answers. This has come back to my split brain attitude. If I'm just too skeptical, I become cynical and I set my ass and do nothing. No, you have to take responsibility for the things that you build. So some people out there are going to build shit and we have to hold them accountable and put pressure on them. But that doesn't mean that we can roll over and see the territory and just say, ah, it's all inevitable is going to be this crazy ass shit. It's going to

Work vs. Jobs

end up being a dark age. He isn't going to be a dark age. I think it's going to be the most productive few decades in the history of human species, and I think it's going to liberate people from work. I think it is going to create, you know, think about it. 200 years ago, the average life expectancy was 35 years old. You and I would be dead today. We're in the 70s and 80s is unbelievable. And some people go all the way up to the hundreds. And I think that that is a massive amount of, like, quality life that we've added to the existence of humanity as a result of science and technology. But, you know, these things are not, you know, sort of

they won't on their own be net good. They'll be net good because we get better as a species at governing ourselves. And so the job is on us collectively as humanity, to not run away from the darkness, confront the risk that is very, very real and still operate from a position of optimism and confidence and hope and connection to humanity and unity and all those things. Like we have to live by that, because otherwise, you know, it's just too easy to be cynical. Now, I hear you. I actually actually agree with what you're saying there. And I think there's one part I'd augment, though, is I wouldn't say we want to get rid of work. I'd say we want to get rid of jobs. Right, exactly. And I think there's a difference between the two, because working is something that brings humans fulfillment, you know. And you see this in children. I always think, like, I'm always fascinated by how you give a child blocks and they start building and they sweep the floor and they put things and they move things around. No one's paying them by the hour. No one's telling them what to do, but they just in their own little brains go, I want to be doing something. And they like seeing the progress of what they're doing. They see the puzzle getting completed, they see the toy slowly forming. They see the colors filling in the picture. You know what I mean? And so I, I feel like that's their work. The difference is, when you make it a job, go clean your room and they're like, ah, damn it, you know what I'm saying? Because now you have a boss telling you to do something you don't want to do. No. And I think to myself, like, I go in a, in a, in a perfect world which we may never get to, but in a perfect world, we find a way for everybody's work, which is their passion to find the other person that finds value in it. Because music has become people's work. And if you think about it, it's crazy, right? People out there, Taylor Swift, billionaire, supremely talented. But if you like rewound time and you said one of the richest people on the

Vision of Abundance

planet is going to be someone who plays a stringed instrument, go back in like the Middle Ages time and be like, yo, you know what I mean? That guy, like busy strumming. They'll be like, you're crazy. You know what I mean? The guy in the town square. Who's playing that little? The richest. But it's because in this day and age that work has been regarded as valuable. Same as playing a sport. Same as working in tech, you know. And that's like my dream world is where it's like beyond the money number. It's like the value of everybody's work comes to fruition because you do have value. If you like knitting, you do like you do have value. If you're a sculptor, you do have value. If you are a poet, you have value for your philosopher, you have value if you're a coder, if you're an architect and engineer, whatever it is you have, that's like my dream world. I actually wonder what yours is like. What's if you could look at if if everything went right, let's say you were now predicting for yourself, you know, you we've managed to survive the wave. We've we've found a way to minimize the risks that come from these small, you know, hostile actors. We've found a way to get governments on board and actually have them understand why they should be involved. And, you know, like, responsible for the for their constituents. Where are we then and when will you say, are we succeeded? We did it. I think that's the real vision is like disconnecting value from jobs, because value is like, you know, everything that you've just described is the experience of being in the physical world and doing something that you're passionate about, that you, that you love, like. And I really believe that there is a moment when, you know, we actually do have abundance and abundance. You know, some people say, well, we have abundance today, right? But it's just not evenly distributed or that it's not enough. And we still want more and more and more and more. And I don't know that with true, true abundance where, you know, because we have a form of abundance today, but energy still costs things.

It's still expensive to travel the universe, to travel the globe. You know, everything is still. we still it's expensive. We have more than we had before, but it is possible to imagine a world like I'm 40 years old. In 40 years time, 2065. It's totally possible to imagine a world of genuine infinite abundance where we do have to wrestle with that existential question of who am I and what I want to do. Yeah, I tell you what I would want to do and that I want to do now. I want to spend more time singing. I joined the gospel choir like 15 years ago for like half a year. I can't sing Jesus, man. Like, I

Global Impact

sound like a strained cat like it. But the feeling that I got inside from being welcomed by this group of people and just being, like, bopping along at the back, I don't think I've ever experienced anything like it. There's just incredible. It's like the most intensely beautiful thing ever, just to be part of a group and, you know, just letting this kind of music come out. So it sounds scary to have to answer that question. But I also I think that if everybody just takes a minute to really meditate on that question, it's a beautiful aspiration and it's within reach. It's within reach. That's what is possible if we really can get it right for everyone and we kind of get obsessed with like, again, we just have this West and think about how many people are earning $2 a day. Think about the 3 billion people on the planet who live by our standards, true, like poverty, lifestyle, um, simply because they don't have access to basic vaccines or clean running water or consistent food. That's like the true aspiration. That's the true vision. And I think, you know, I think that's within reach in the next 20 to 40 years. It's really eradicating that kind of suffering. Um, so some of those test cases, you talk about some of the more inspiring ones I've seen, I'm sure you have as well. Uh, when I was in India, I got to travel with a group there who was using AI to predict which houses would be most devastated by a flood or a storm, and so they would get people to remove their belongings. Because most people are one devastation away from losing everything they own. And so they would use AI to track where these things are and where they're going to be. And they could even tell you in a heat wave who should leave their house so that they don't die. And you look at programs in Kenya where they've used AI to help farmers not lose all of their crops. They could tell them now and they use it on like a flip phone. They've told Kenyan farmers, hey, here's your phone, you've got your own little AI and it'll just tell you when to plant, when not to plant, when to not waste your seeds, when to. And it's increase their output, you know, like 90% where before it was like a gamble and they were losing, you know, in 1 in 1 harvest they could lose everything, right? All of a sudden they have it. And so I really do like what you're saying, because on the one hand, there's always the risk of losing something. On the other hand, there's the opportunity of gaining sort of everything, and the balance is where we have to find it. Well, and the proliferation of those technologies are so much more subtle. Like, it sounds like it's just this binary thing of getting access or not getting access, but it's so it's so nuanced. You can't even tell how much good it's doing. Like, I was reading this stat the other day that like three years ago, um, 10% of households in Pakistan, um,

Community Lessons

had solar panels on their roofs, which meant that some very large percentage were still burning biomass inside of the home and getting all of the kind of breathing issues and everything else that comes with that. The cost of solar has come down so dramatically just in the last three years that within, I think it was 18 months, the number of personal like consumer households that adopted full on solar on their roofs went from 10% to like 55% in 18 months. Crazy. Just because it suddenly became affordable and it crossed that line,

suddenly everybody has. It's insane. You know, near free energy, which obviously means they have access to phones and laptops and connection to the digital world. And they're able to, you know, do all the things. So I think that it's easy to overlook what is already happening around us. All the good that is already happening around us all the time and how fast it's happening. Um, it's it's too easy to get trapped in the cynical world, and I think it's a choice not to. There's a choice to be aware of it and hold it and take it seriously, but not be like, owned by it. Before I let you go, there's one question I have to ask you, and it just you just brought it up when you were talking about, like, Pakistan and these places, one of one of the programs you started, um, in the UK, was started right after nine over 11 and was basically a helpline in and around. It was like Muslim people who were being targeted. It was this was just rampant Islamophobia after nine over 11. Right. And you stepped in with a few people and you're like, stop this program, because I want, like, Muslims to just have a helpline, you know, talk about whatever they need to talk about. And it's still running till this day. I think it is the number one, if I'm not mistaken, like the, the largest, like Muslim specific helpline. And I couldn't help but wonder how that shapes how you think about AI. And what I mean by that is tech has often had one type of face attached to it, you know, and not in a bad way. I'm not like villain izing any. It's just like, yeah, that's where the thing was. But as tech slowly starts to evolve, you know, you see like India hopping up as, like a as, like a powerhouse. You know, I remember when I was traveling there, you just go and you're like, wow, this is like a new Silicon Valley. And what's coming out of it is different and impressive for a whole host of different reasons. You know, Nigeria has a whole different type of Silicon Valley, Kenya, as I said, South Africa. You look at parts of the Middle East as well. And and I wondered like, how does that shape how you think about AI? Because we often hear stories about like, oh, AI is just going to reinforce bias season and AI is just going to be like, oh, you think racism was bad before? Imagine racism in The Terminator. Imagine if Arnold Schwarzenegger was like, I'll be back, nigga, you know what I mean? It's way worse now. Like, so now when you think about like that tech and you're in it and you know what it's like to be in a group that is ostracized. How do you think about tackling that? How does that shape what you're trying to design? We created Muslim Youth Helpline as a secular, non-denominational, um nonracial, um group of young people led by young people, for young people. So it wasn't about being religious. It had Sunni and Shia had Pakistani, Somali, English, white, everything you can think of. And we were all like 19, 20, 21 years old. The first CEO was a woman, a muslim woman.

Closing Thoughts

And so my response to that feeling of being threatened, basically post 911 with rising Islamophobia and being accused of being terrorist was to create community. And that community taught me everything, taught me resilience, respect, you know, empathy for one another. And that's the simple act of listening to people just being on the end of the phone and using a little bit of faith and culturally sensitive language in nonviolent communication, just making people feel heard and understood was this superpower. You're not telling them what to do with their life. It's non-judgmental and unidirectional. You're just making them feel heard and understood. And that has always stayed with me. It's been a very important part of my inspiration. And, you know, it speaks to a lot of what I've been doing now, especially with my previous company inflection and pie and stuff. You know, pie was really like a very gentle, kind, supportive, you know, listening AI and I was then I remember using it. Yeah, yeah. And we became part of Microsoft now. And so I think that's what makes life worth living. And that's why I still try to, to do if I can today. Yeah. Oh man. Well, you know I

appreciate you. Thanks for taking the time today. Uh, thank you for writing the book. I'll recommend everybody read it, like, as long as they can. And because it's, it's it's, um, it's a very human look at a very technical problem. And I think that's sometimes what AI is missing. I talked to a lot of engineers who don't understand the human side, and I talk to a lot of humans who don't understand, like, the engineer side of it. Um, but yeah, man, thank you for taking the time. Thank you for joining us. And I hope we have this conversation in like ten years. Just me and my baby, little AI, and we're just going to be talking to you about it. Just be like, you know, what? Do you want to ask my stuff out here? Ask Uncle Mustafa a question. I'll be back. This has been amazing, man. Thank you so much. Thank you. Really appreciate it. Thank you. Shit. I think you split my brain in 15 pieces. Oh, man. That's amazing. Thank you very much, man. Thank you so much for watching the episode. If you enjoyed it, pass it on to a friend. If you didn't enjoy it, pass it on to your friend. Still let them suffer for a change. And don't forget to engage with us in the comments. If you want to suggest a guest, maybe there's questions you want. Maybe there's ideas that you have. You can chat to us, we will read through the comments and we'll get into it. Either way, I appreciate you taking the time. Thank you for hanging out with us on What Now? And remember we do this for you, so we would like to hear from you till next time. Thank you.