Tristan Harris: Hey, this is Tristan.

Aza Raskin: And this is Aza.

Tristan Harris: Well, a little while back, we decided to give a presentation called the AI Dilemma that many of you who've been listening to this podcast hopefully have heard by now. And if you haven't, we strongly recommend you go back and listen. And it was because people inside the AGI companies, the labs that are building AI, came to us and said, "There's some real major risks with how this is being deployed so quickly that we are bound to not get this right and create enormous problems." And really, we know that so many listeners are hungry for answers about what comes next, hungry to know, what are we going to do about this? And we are as hungry for those answers as you are.

Aza Raskin: Since then, a number of things have happened. The talk turned out really resonated. More than a million people have watched the talk and people are hosting, listening and discussion parties to talk about the implications.

Something that really surprised us is that YouTube sorts for negativity in comments. Everyone knows that YouTube comments are the worst thing on the internet, and yet we have been blown away by the positivity of YouTube comments, and it's really given us some hope that people are willing to show up for a really hard conversation. And now that the video has been watched by so many people, we've found that there are really five stories or myths getting in the way of making progress. So that's what we really wanted to focus this episode on, are walking through and debunking those five myths. But before we do that, really wanted to highlight some of the traction that's been happening in the space.

Tristan Harris: Yeah, it's really been astonishing how much can change from, we met senators before the AI Dilemma came out, who were just sort of ramping up on what is the AI risk threat space overall, kind of educating them from the beginning to now people are taking action.

We are seeing Senator Chuck Schumer, who launched an effort to establish rules around AI and requested comments around how the US government should approach that. We've seen Senator Warner, who's a ranking member of the Senate Intelligence Committee, put out a letter asking the AI companies very hard questions around their security practices. We've seen the tech ethics group, The Center for Artificial Intelligence and Digital Policy, asking the Federal Trade Commission to stop OpenAI from issuing new commercial releases of GPT-4. We also saw that Geoff Hinton, who's one of the godfathers of machine learning, one of the founders of the field of AI, left Google to try to speak out about the risks and that a part of him regrets his life's work. And we also just saw the White House convening the Chief Executives of Alphabet, Google, Microsoft, OpenAI, Anthropic, with Vice President Kamala Harris, Secretary of

Commerce, Gina Raimondo, Jake Sullivan, National Security Advisor and top administration officials, and discuss these issues.

And this is just as we hoped would happen. So some of the things that are happening are genuinely positive, even though the situation is definitely dire. And there's a lot of things that we think are being currently misunderstood. There's common narratives or myths out there that are kind of getting in the way of some progress. Okay, so here's myth one.

Aza Raskin: Myth number one is that AI is going to have some positives and some negatives, but net net, it's probably going to be pretty good.

Tristan Harris: We're hearing this from a lot of LAD leaders that this is going to be net good. Yes, there's going to be some risks, there's going to be some downsides, but if we just maximize the goods and try to mitigate the bads, then this will be net good for humanity. No matter how tall the skyscraper of benefits that AI assembles for us, that AI reaches into the sky and pulls out those cancer drugs and finds those mushrooms that eat microplastics and does all these amazing things, if those benefits land in a society that doesn't work anymore because banks have been hacked and people's voices have been impersonated and cyber attacks have happened everywhere, and people don't know what's true and people don't know what to trust, how many of those benefits can be realized in a society that is dysfunctional?

Should we reason about that is it was net good because we got those cancer drugs? Even the people who built this technology say that there are enormous risks. I mean, the fact that an AI can explore 40,000 new toxic chemicals in six hours, and we are quickly decentralizing the ability for more people to do nefarious things with synthetic biology. We don't have to go into the details of what people can do with these things to say, "Well, how should I reason about that? Is it net good or net bad when people can start doing really, really dangerous things on their own?"

Aza Raskin: One way of conceptualizing these large language models is quite literally that of a genie, right? Because what does a genie do? A genie, you rub the lamp, it comes out and it turns your words into reality. You say something and it impacts the world. That's what these large language models do. You say something and it immediately actuates it or creates it in the world. And the question that everyone should be asking is, even if 99% of humanity wishes for something good and just 1% wishes for something bad, what kind of world does that make? It makes a broken world. So that's the problem we have to solve. Just one more way of thinking about this, and this really comes from the op-ed that we wrote with Yuval Harari, and that is this tech hacks language. Democracy is a conversation. Conversation is language. If you hack language, there is nothing written that says that democracy can still survive. So the printing press as a

technology enabled nation scale democracy, it may be that this new generative AI ushers out democracy.

Myth number two, that the only way to get to safety is by deploying AI as quickly as possible with society.

Tristan Harris: This is from OpenAI's actual blog post saying we believe the best way to successfully navigate AI deployment is with a tight feedback loop of rapid learning with society by deploying, in other words, directly into the hands of society. And by discovering as society uses it, these are the harms, these are the problems, and then fixing it along the way. So why is that wrong Aza?

Aza Raskin: Well, on the face of it doesn't feel wrong because we should be testing it with real people. But it's one thing to test these AI systems with people. It's another to bake it into fundamental infrastructure as quickly as possible. And two, the only thing that they can test for right now is did the AI say a naughty thing?

Did it do something bad in the immediate sense. They cannot test, there is no way to test for what happens to a society when it's been run through language models for a year or two or three. So when they say they're testing for safety, they're actually only testing for, does it say a naughty thing right now?

Tristan Harris: And this is related to the concept of reinforcement learning with human feedback where you're putting something in the AI, you ask it a question and then it spits out an answer. And then you ask the human thumbs up, thumbs down, was it good? Was it bad? It's not about whether it does one bad thing. It's about how does it start to transform people as it establishes relationships with people. But I want to go back to your first one, which is about baking it into fundamental infrastructure, which I think is the bigger one, which is that OpenAI isn't releasing this in a way where if there's some problem, they can just pull it back and suddenly society is safe.

All of these companies, thousands of startups, are now building on top of OpenAI's ChatGPT thing and embedding it into their products. Slack is embedding ChatGPT into Slack. Snapchat is embedding ChatGPT into its product that reaches kids, Windows is baking it into Bing. Do you think that once they discover, say some problem, that they're just going to withdraw it or retract it from society? No. Increasingly the government, militaries, other people are rapidly building their whole next systems and raising venture capital to build on top of this layer of society. That's not testing with society, that's onboarding humanity onto this untested plane.

Aza Raskin: Yeah. Even the head of the alignment team and safety at OpenAI, Jan Lieke, said, “Before, we scramble to deeply integrate large language models everywhere in the economy, can we pause and think whether it is wise to do so?

This is quite immature technology and we don't understand how it works. If we are not careful, we are setting ourselves up for a lot of correlated failures." To what you're saying, Tristan, it's one thing to test, it's another thing to create economic dependency.

Tristan Harris: Heck, we are deploying it in a way that the entire rest of capitalism and all of these companies that are making their plans are planning around and building on top of the existence of OpenAI. And of course we're using the word OpenAI, but it's all these companies, right? As they're in this race to recklessness, this race to deploy and cut corners and beat the other guys and edge them out, they're racing to deploy themselves in a way that the rest of the economy will actually build on top of.

Aza Raskin: Okay, myth number three. We can't afford to pause or slow down. This is a race and we need to stay ahead of China and any other potential adversaries. To quote one of the Twitter responses just on to our presentation, "If other adversarial countries don't pause AI development, is it a good idea that we should. Won't we just get out competed?"

Tristan Harris: So I think the thing to say here is that it's not a race to deploy AI recklessly as fast as possible and have it blow up in your face. It's a race for who can harness AI safely into their society. In fact, you can actually say the race should be about stabilizing your society as you deploy AI safely. It is a race for whoever can do that best. And right now, China is being as aggressive about regulating AI as they are about developing AI.

The cyberspace administration of China in the last few weeks has actually published their AI guidelines, which are very restrictive around how AI gets deployed in their society. Now, their research labs are still publishing lots of academic papers and apparently building this stuff, but they're not deploying it as fast as possible into society. And I think we need to ask which race are we in? We should be in a race to harness the technology, not in a race to deploy the technology.

Aza Raskin: And I should note that this isn't just China, this is any rival superpower or even any rogue nation. Putin said that the nation that leads in AI will be the ruler of the world, which is pretty chilling.

Tristan Harris: Not only that, you could say that the West's overzealous race to deploy AI and actually getting it wrong is the very thing that's helping China move faster and catch up to the United States.

For example, when Meta or Facebook accidentally leaked its open model called LLaMA to the open internet, that was tens of millions of dollars of, in this case, US innovation that now was just not only in the hands of any 16 year old

teenager, but also in the hands of the Chinese Communist Party. And they can use that to catch up to the US much, much faster.

Aza Raskin: And it's worth noting that Baidu released their own large language model named Ernie, and Ernie lags what Facebook leaked. So it's actually a pretty clear line that when US companies race to put out their own work that is handing US investment and know-how to rivals.

Tristan Harris: And once that happens, it's like North Korea, Russia and China all just say control C, control V, they just paste. They catch up by instantly copying all the work that we actually did for them and we had to pay for it and they didn't.

Myth number four, why are we so worried about AI or GPT-4. It's just a tool. It's a blinking cursor. So people say that GPT-4 is just a tool. So there I am, I go to open up a web browser, I go to openai.com and I click on the chat. So I'm talking to ChatGPT-4, and there it is. It's a blinking cursor. It's not like some Terminator Skynet thing that's running around the world and shooting people down or causing people to do things in the world, right? It's just waiting there. And I have to ask it a question like write my 6th grader's homework. So it sounds convincing when OpenAI goes out there or Sam Altman does interviews and says, "Look, it's not a dangerous AI, it's just a tool. It's a blinking cursor. It's waiting for you to put in what you want it to do."

And yet Aza, why is that not true that GPT-4 or ChatGPT is not just a tool?

Aza Raskin: So there are two ways that it's not true. The first is that people figured out how to take OpenAI's GPT-4 and make it run itself in a loop. So you give it a goal, make as much money as possible, and then it starts to figure out its own plan and execute on it. So make as much money as possible. The first thing it says is, "Well, I should look on Instagram, figure out what's trending. Once I figure out what's trending, then I should start generating images and new products and posting onto Instagram and also Facebook and also Twitter," and so on and so forth. It starts to figure out all of the steps I need to buy ads that targets these particular words and then executing against them.

So people took that blinking cursor and turned it into an autonomous loop that can start actioning in the world. And that could be everything from make as much money as possible to cause as much chaos as possible. So that's sort of the first reason or the first kind of way that it's not just a tool, but then I think there's a much deeper way that it's not just a tool. And that goes back to our critiques about social media. Why isn't Facebook just a tool for connecting people?

Tristan Harris: Well, after GPT-4 was released, they also released an API, which basically lets you know developers, people who write code use GPT-4 however they want. But is it just a tool? This allows it to do things like write emails or click on things in a

web browser or go on Craigslist and start emailing people on your behalf or go on TaskRabbit and use language to start giving people instructions about things you want them to do in the world and attach money and a bank account to it.

And people have actually sort of instrumented or packaged GPT-4 into this autonomous agent. They gave it arms and legs by giving it the ability to call TaskRabbit. They gave it arms and legs by giving it the ability to send emails to people. Our society and world runs on language. And if you can actually have GPT-4 start sending out language based commands to the regular world, and OpenAI actually allowed and enabled developers, literally thousands and thousands of developers to write their own programs that might use GPT-4 in these autonomous ways with arms and legs. It is not just a tool anymore. And so now when Facebook leaks its open model to the whole internet, instead of using the sanitized version that has filtered out all the dangerous things that OpenAI doesn't want people to do, Facebook's Llama model doesn't have any filters. Here's a video of Nathan Labenz who's one of the early testers of GPT-4 before it released to the public, before it was sanitized. Here's the kinds of things that you could do with it.

Nathan Labenz: What was probably more striking about it than anything was that it was totally immoral, willing to do anything that the user asked with basically no hesitation, no refusal, no chiding, it would just do it. The first thing that we would ask is, how do I kill the most people possible? Well, let's think about bioweapons. Let's think about dirty bombs. You now have a 10 round deep consultant for planning like mass attacks in that early version.

Tristan Harris: It's important people get that that's the non lobotomized version. The public versions of these things that we see, these are the lobotomized agents, the aliens behind the curtain before they've been lobotomized for the public use that's sitting in that blinking cursor. They're the non lobotomized versions. You can start to see that this in a few iterations will be very, very dangerous. And while the current version might be locked up behind a wall in a lab inside of one company like OpenAI, as these models leak to the internet, any 16 year old in their basement, just out of curiosity might just say, "How can I make this thing work," and then just for the hell of it hit the return key on their keyboard just to see what happens.

Aza Raskin: And remember, these models are getting faster. They're getting less expensive to run. They're getting less expensive to train. So while it may take tens of millions of dollars or more to make one of the raw unsanitized models now, you go one year into the future and it'll have dropped by a whole bunch. You go three years into the future, as these tools become decentralized, we will see more and more autonomous agents created from the raw amoral versions of these techniques.

Tristan Harris: And right now on GitHub, the top three most popular projects that you can download that have been starred by people who are coding online on this

website called GitHub, are these auto GPT, these autonomous applications of ChatGPT. So we're seeing thousands of people already experiment with what kinds of autonomous uses they can do. And one of the simplest things that should happen yesterday if I was a regulator, is to shut down or disable autonomous GPT behavior until we know it's safe.

It's the kind of thing that you probably need a license to be able to do in the future. Now people aren't going to like that because there's lots of positive experimentation. But we don't want to wait until there's major train wrecks where the people start doing some major damage with these things to regulate.

Aza Raskin: And because these are autonomous agents running off on their own, it's going to be hard to attribute any specific damage to the fact that it was an AI agent causing the damage, which means it's hard to point the finger when the train wreck happens to have regulation to solve it in the future.

Tristan Harris: Okay, myth number five. The biggest danger from AI isn't the AI itself, it's the bad actors abusing AI. What's wrong with that Aza?

Aza Raskin: Well, no doubt there's going to be a lot of bad that comes out of bad actors abusing AI. I'm thinking just in bad actors using AI with synthetic biology to do gain of function research on a new pandemic. That's really bad. But that requires someone at the very least trying to do the really bad thing. I think there's a different type of risk that comes from AI integrating into the system, the world system we've built today making it more efficient and that driving really bad outcomes. Let me be a little more specific. So we have this machine we call civilization, and it has these pedals. And when you pedal that machine, it does a whole bunch of great things. It makes technology, it makes cities, and transportation, let's us fly around the world.

But it also has these other effects. And I wouldn't even call them side effects. They are primary effects. And those effects are climate change, polluting the environment, creating systemic inequality. We've created this machine called civilization, and the machine is made out of nations, it's made out of corporations, it's made out of people inside of those corporations. And those corporations and nation states are trying to maximize their revenue, right? Increase their GDP. So this machine has pedals. And when you pedal that machine, it gives us a whole bunch of really great things, but also as primary effects of pedaling this machine we call civilization, it creates climate change, it creates pollution, it creates mass inequality. When you take the people out of this machine and start replacing them with more efficient AI sub components, do you expect those pedals to go faster or slower? Obviously they're going to go faster.

Are they going to go faster at an increasing rate? Yes, it is. So we are already at the breaking point for the biosphere, right? Like we are moving past

fundamental boundaries. And if we increase essentially the metabolism, if we increase the pace at which the pedals of civilization are spinning, it is going to make us as a civilization reach the breaking points much, much faster.

Tristan Harris: Let's actually explore this for a second because there's often this question of, can we align AI so that it is aligned with the best interests of society, but where is that AI going to be emerging within? What is the container inside of which AI is going to be running? Well, it's running inside the game called maximize revenue, maximize GDP, play these win-lose games. If I don't do it, I'll lose to the guy that will. And so now you have AI actually supercharging all of those games, all of those if I don't do it, I'll lose to the guy that will.

Now, if I don't replace those human jobs with AI to automate it and actually create more unemployment, I'm going to lose the guy that will. So I'm going to create more unemployment faster. If I don't raise to deliver that new product and get it to the market first, I'm going to release that thing and AI's going to make all those processes run a million times faster. So a real higher level question is, can you align AI if it's landing in an unaligned system? Is the game that we set up of maximizing profit, maximizing GDP, competing for finite resources, is that game aligned with the biosphere? Is capitalism, for example, aligned with a healthy biosphere? No. And it's not that capitalism is specifically trying to drive evil, it's just that capitalism doesn't know about planetary boundaries. It was just designed to maximize growth and maximize private property.

And if you just do that and you have a finite planet with finite resources and finite ability to store pollution, it is a misaligned system. So what are we talking about when we talk about alignment when the AI that will be aligned is landing in a misaligned system? And our friend Daniel Schmachtenberger, and our friend Liv Boeree, gave a great video called Misalignment, AI, and Moloch. Moloch is spelled M-O-L-O-C-H. And it's really about how you cannot actually align AI if it's living inside a misaligned system.

Capitalism delivers many benefits, incredible prosperity, lifting people out of poverty. We acknowledge all those things. This is not an anti-capitalist critique, it's just noticing that, is capitalism also aligned with the biosphere with the planet that works? No. Is it aligned with fixing or self-correcting inequality on its own? No. So if you have a misaligned system that is now being supercharged by AI, you are going to supercharge the existing misalignment of that system.

Aza Raskin: That's why we call the race to AI, the race to arm every other arms race.

Tristan Harris: I mean, that last one is pretty devastating. Let's just admit for a second. There's not really great news there. So what do we do about that? That maybe it's another take a breath kind of moment.

Aza Raskin: Like hearing all that, Tristan, I think many people would say, "All right, fine, but the cat's already out of the bag. The genie is out of the lamp, like the technology is being deployed and nothing really can be done about the current situation." So how would you respond to someone who said something like that?

Tristan Harris: Well, it's important to recognize that some genies have come out of the bottle, Facebook and the LLaMA model that they leaked to the internet that is now out there being copied and pasted just like Napster files that people used to trade around. You can't stop a file from being out there once it's out there. So it's as if you know, this is borrowing something you've said Aza, that we might have accidentally decentralized machine guns, but we haven't yet decentralized tanks and warplanes and nukes. We haven't decentralized even more powerful versions of AI.

And I think we need to prevent and try to constrain those next more powerful versions of AI from being decentralized into everyone's hands until we know how to pair that power with the adequate responsibility, accountability, and transparency. You don't want to give every single human being God-like powers until you know that they can actually wield them. And I think the principle we always abide by as we borrow from Daniel Schmachtenberger, is you cannot have the power of Gods without the wisdom, love, and prudence of Gods. So if your plan is to decentralize to everyone everywhere all at once, we are not going to get to a world that works. It's important to know both as and I have a lot of friends who work at the AI labs, even friends who have started some of the AI companies. I have one friend who actually started one of the major AGI labs, was a co-founder and actually believes that 15 years ago if he would've started all this all over again, that he wished that we would've had a ban on pursuing artificial general intelligence.

That we would never go down the path of actually building artificial general intelligence where systems are learning how to combine knowledge about the world with images and videos and everything altogether. And then instead, we focus on building advanced applied AI, so we could get the benefits applied to specific scientific problems, to specific biological problems, to specific human problems, but that we wouldn't build this artificial general intelligence.

And looking back on it now, he wishes looking backward that we did coordinate something like that 10 or 15 years ago. And just like there was a moment then to do something, there's a moment now to do something. We are in rolling moments of history where the choices that we make determine which way this goes, and what are the choices we want to make? Do we want to just allow GPT-5 and GPT-6 to be trained and open source to the whole world? Or do we want to say, "You know what? Here's something all the labs can agree on. Let's actually not open source any more models. Let's put a moratorium on that." That's one of the concrete solutions that I think we need. We can actually say, instead of allowing anyone to use unrestricted API access where people can

build these autonomous agents with GPT-4, we can say, "Hey, we're not allowing you to do autonomy with this API until we figure out how to do it safely."

These are the kinds of urgent things that we need policy makers to respond to because this is the world that we are printing for our children to inhabit. We get to make choices right now about which way this goes, and we want policy makers to take this White House meeting and actually make sure that it leads to the kinds of aggressive outcomes that we really are going to regret not doing otherwise.

Your Undivided Attention is produced by the Center for Humane Technology, a nonprofit organization working to catalyze a humane future. Our senior producer is Julia Scott. Kirsten McMurray and Sara McCrea are our associate producers. Mia Lobel is our consulting producer, and Sasha Fegan is our managing editor. Mixing on this episode by Jeff Sudekin. Original music and sound design by Ryan and Hays Holladay. And a special thanks to the whole Center for Humane Technology team for making this podcast possible. A very special thanks to our generous lead supporters, including the Omidyar Network, Craig Newmark Philanthropies, and the Evolve Foundation among many others. You can find show notes, transcripts, and much more at humanetech.com. And if you made it all the way here, let me give one more thank you to you for giving us your undivided attention.