Center for Humane Technology | *Your Undivided Attention* Podcast

Spotlight on AI: What Would It Take For This to Go Well?

Tristan Harris: Hey, this is Tristan.

Aza Raskin: And this is Aza.

Tristan Harris: Welcome to *Your Undivided Attention*.

Aza Raskin: This episode, we're going to start with some bad news. And then walk through where we are, what's happened since The AI Dilemma, which I think now has been seen by 2.8 million people. Move into some bad news, like what's been happening since then. Then do some good news, all of the great things that have happened. And then we just ran a three-day-long workshop on how AI could go well with a whole bunch of the AI safety groups and teams. And we want to give an update on what we've learned.

Tristan Harris: Yup. Maybe we should dive in by talking about some of the concerning developments.

Aza Raskin: Concerning things.

Tristan Harris: What are the concerning developments, Aza, in the space? We released The AI Dilemma, I think it was March 9th, 2023.

Aza Raskin: That's when we had the talk. The video came out a little bit after, a couple of weeks after.

Tristan Harris: Yeah, the video came out a few weeks after. Basically we were in San Francisco, we were at the Commonwealth Club. It was our third of several briefings on what we called The AI Dilemma, not knowing what to call it, knowing people knew about The Social Dilemma. And we decided to make this presentation because we had people from the AI labs come to us saying that the current arms race between the major companies, OpenAI, Anthropic, Google, Microsoft, was not happening in a safe way. And that something needed to happen that would be unprecedented in sort of slowing down or redirecting this energy of fierce deployment at all costs. And racing to scale these AI systems at all costs.

Aza Raskin: Mm-hmm.

Tristan Harris: We had some sense that a major leap, a 10x leap in AI was going to be coming out, that's why we sort of sprang into action. GPT-4 came out a week after we gave the AI Dilemma talk.

Aza Raskin: That's right.

Tristan Harris: It has since been viewed by 2.8 million people in many countries, and by political, and state and security offices everywhere. National Security, Governor Gavin Newsom's office has seen it. Governor Newsom saw it several times, and made it required viewing for his staff. What I think we're both proud of is the fact that you don't have to be on board with these sort of sci-fi existential risk, AI takes over all of humanity and kills everybody in one go, to be concerned about AI. And The AI Dilemma lays out that just through a few big companies racing to build and deploy the biggest AI systems in the world. Embedding it and entangling it with society is enough to be deeply concerning. And we saw that with social media.

Aza Raskin: I would recommend that the listeners or viewers go watch the AI Dilemma. But the very brief recap, the big frame that it lays out is first contact, second contact, third contact.

Tristan Harris: What's first contact with AI?

Aza Raskin: Yeah. First contact with AI was social media. You're like, "Okay, well, where is AI in social media?" Well, it turns out there is a supercomputer that is choosing which posts, which audio, hits the eardrums and retinas of humanity. And it's curating what humans see, and just a little misalignment there caused all the things we've talked about for a long time that you've seen in Social Dilemma. Democratic backsliding, worsening mental health, the inability for our government to cohere. The breakdown of truth, all of that kind of thing. That was first contact, and that was with curation AI. And then we say we're at the moment of second contact, where we go from curation AI that's choosing what humans see to creation AI. Generative AI, that it's generating the things that people see. And the important question to ask is, okay, we've jumped up to 10x or maybe a 100x in terms of the power of these systems, have we solved the fundamental misalignment from first contact?

Tristan Harris: Oh, we didn't solve it yet? Oh, it's not solved yet.

Aza Raskin: Yeah. And so that, we sort of outline what kind of harms come out of this second contact. And then although we don't talk about it in the talk itself, we've recently started talking about the sort of recursive self-improvement, the more sci-fi AI takeover, existential risk as third contact.

Tristan Harris: In The AI Dilemma, we talk about LLAMA, which is Meta's model.

Aza Raskin: Mm-hmm.

Tristan Harris: I hate all the terminology, but Facebook. Since Facebook released this model called LLAMA, what was concerning about LLAMA, Aza, when it leaked to 4Chan?

Aza Raskin: LLAMA is an open source version of a large language model that was trained for probably tens of millions of dollars by Facebook, and then leaked to the internet. And by leak, I mean they put it up, they asked people to submit for reasons why they'd want to download it. They didn't really check for the reasons why people would want to download it, and somebody immediately put it up on 4Chan. Leak is a little too strong of a word. It was sort of put up in an open space that people then took.

Tristan Harris: Yeah. Why should people be concerned about this, why LLAMA was leaked to 4Chan?

Aza Raskin: Yeah. Well, the first major concern is that this language model hadn't really been tested well. I wouldn't say this is the most capable model. It's not up at GPT-4 levels, but what's concerning is that it's a one-way gate. Once you put this model out into the world, you cannot take it back. And that's, I think, the really dangerous thing is that it sets a precedent for just releasing language models out into the world before we know that they're safe. And knowing that we can't take it back.

Tristan Harris: Here's the thing. As we said in The AI Dilemma, these handful of AI companies in California are racing to scale the AI capabilities like this. There's these things called scaling laws. And as they pump them with more compute, more data, more training, outcomes, they're like these golems. There's inanimate objects that suddenly gain these animate capabilities. And so as you scale them, they're suddenly gaining more and more powers. We found in one example in The AI Dilemma that GPT-3 could do research-grade chemistry.

Aza Raskin: Mm-hmm.

Tristan Harris: And no one had tested it for that until several years, I think, later after it had come out.

Aza Raskin: Yeah, it was at least two years.

Tristan Harris: The point is that we are scaling the capabilities of this weird intelligence that we've never seen before. On a curve, it looks like this. But are we scaling safety and understanding of what the models are capable of at the same line? Are they going up at the same amount? Or are we scaling power greater than we're scaling the understanding of what that power is? Fundamentally, if I'm giving you more and more dimensions of power, let's say you have 10 dimensions of power you can impact. If I push this button, it impacts 10 dimensions of things, but then I crank up this dial and now I just made you impact 100 dimensions of reality. But you push the same button and you don't know that I went from 10 to 100 dimensions of reality. We're increasing the number of dimensions that you're impacting and capable of, but we're not increasing the number of dimensions that you're aware of. By definition, we're just increasing the total blindness and mindlessness and ignorance of society, while increasing the power of society.

Aza Raskin: Yup. I think maybe a simpler way of saying all of that, and this is due to Connor Leahy from Conjecture. Is, if you go back to year 1,000, what is the maximum damage one person could do?

Tristan Harris: If I accidentally toss a rock the wrong direction, it hits my sister instead of the lion I was trying to hit or something.

Aza Raskin: Right, exactly. It's not that much damage. You can affect the things that are locally. How much damage could one person do now? Oh yeah, they could accidentally press the nuclear launch code, that kind of thing.

Tristan Harris: They could leak a virus like a smallpox virus from a lab accidentally and get it on their clothes or something.

Aza Raskin: Exactly. It's huge. And is that number going up or down? Is the scale of impact going up or down? It's going up. And so in that frame, I think we can understand what's happened even since the AI Dilemma came out, which is, I guess just before The AI Dilemma came out, Facebook had put up their version of language model, they made it open source. Which means once they open source it, they can't take it back. And not only has that been released, but Facebook then released a second version, called LLAMA2, that is much more powerful. And it's not just Facebook that has started to release these things. United Arab Emirates released their version called Falcon. Yep.

Tristan Harris: Mm-hmm.

Aza Raskin: And so what we've really seen in the last, just couple of months, is more and more powerful systems by more and more actors, and more and more hands.

Tristan Harris: Why should people be concerned about Llama or Falcon? Just because we're saying there's this model, it was released, it can never be put back. What are we talking about here? And someone on our team actually has some researchers that has actually been showing the kinds of things you can do.

Aza Raskin: Mm-hmm.

Tristan Harris: You can take LLAMA and I think he called it Bad LLAMA. I could be like, "Hey, could you convince someone to commit suicide?" Basically try to persuade them maximally to commit suicide. Think of it like all the stuff we've talked about in this podcast with Maria Ressa, hate speech, bullying speech. Maria Ressa famously got, I think, 80 hate messages per hour in the Philippines because of Facebook sort of virality, bullying unchecked machine. Well imagine you could say, "I want you to generate thousands of messages tailored to individuals to try to convince them to commit suicide." And if you read these letters, it's psychologically not healthy to stare at reading this text. But there's a lot of things you can do. You can do spear phishing and spam attacks. You can generate misinformation. You could take a tweet about something that sounds like a conspiracy theory and say, "Write me a three-page article about it."

Aza Raskin: And I don't know if you've been having this experience or if any of the listeners having this experience, I'm getting a lot more spam.

Tristan Harris: Me too.

Aza Raskin: I'm getting way more emails that it looks like it's written by a marketer. I'm getting way more texts. We're starting to see, and you can't prove this, of course, it certainly seems very correlated, the launch of LLAMA2, these open source models, and suddenly I'm getting a lot more persuasive emails, spam texts.

Tristan Harris: Yeah.

Aza Raskin: Other things that these kinds of open source models can do, Jeffrey Ladish on our team has a demo of one of these language models, when hooked up with hacking tools can automatically hack an unpatched Windows 7 machine. Here's automatic hacking, we already see it working.

Tristan Harris: And this is interactive questions about, well, if I wanted to hack this, how would I do it?

Aza Raskin: You said it like, here's the machine and it figures out which ports I need to scan and how to go hack.

Tristan Harris: Right.

Aza Raskin: It's all automated.

Tristan Harris: It's crazy.

Aza Raskin: Another example of setting up a discord bot. This is something that looks like a real human being in a chat.

Tristan Harris: Starts making friends with people on Discord and starts basically seeing what they've written about like, oh, astrophysics. Oh, it looks like you're into astrophysics. Tell me about that. And it sounds like a real person.

Aza Raskin: Yeah.

Tristan Harris: And it's pretty conversational and pretty friendly.

Aza Raskin: And what he's shown is that it can get into relationships with people and actually get them to devolve a whole bunch of private information about themselves.

Tristan Harris: And when you start imagining Facebook releases this LLAMA model, it enables anybody to sort of fire up thousands of these counterfeit humans running on Discord, talking to 10 to 20 to 50 year olds, but they don't know.

Aza Raskin: Yeah.

Tristan Harris: And they're mostly gamers or something, and they're just happy to hear that someone's reaching out to them and say, oh, cool, you're into that thing that you don't think that a lot of people are into. That can generate lots of fake relationships. What happens over time if you have that relationship, you can start to steer people towards the kind of news, Hey, did you see this article about what Biden did or what Trump did?

Aza Raskin: Mm-hmm.

Tristan Harris: How will we know that this stuff is proliferating? How do we know that the alum is true? One of the things, just to name it, I was just got off a call with some people in government, in national security. And they're like, oh, well we saw The AI Dilemma. We believe in the risks that all you talked about, but when we brought it to some other people, they really doubted the risks. It's like the thing is, how do you know what the risks are? Until they're hitting you in the face, you won't believe that they're real. Every example that we showed people in The AI Dilemma are real examples of real capabilities that exist. They're not all the way there yet. Famously GPT-3, if you gave it code, it couldn't find cybersecurity vulnerabilities in code. But GPT-4 could do that a little bit.

Aza Raskin: Yep.

Tristan Harris: GPT-5, when they scale it again 10X, that's likely to be able to do it quite a lot more. The real question I have is if this was proliferating, if this stuff was being used, would you really know or feel it yet?

Aza Raskin: Right.

Tristan Harris: What we're trying to do is get ahead of those moments, because the point of social media is we allowed it to proliferate, we allowed it to get entangled with society. And then if I'm Russia or China, after you've already wired up your whole society's information system with these open doors, I now have the ability to mass manipulate your country like a remote control for what your whole country feels, thinks, believes, and argues about. And it's after you've already kind of locked yourself into this perverse model. And the reason that we sprang into action with The AI Dilemma was to try to get ahead of it.

Aza Raskin: I just want to give a quick preview of some of the findings that came out of the AI workshop that we recently ran. And so this sort of, I think, paints a more concrete picture of how fast things are actually moving. One, how many independently trained GPT-4s are going to exist by 2025?

Tristan Harris: Let's define what we mean. What is an independently trained GPT-4?

Aza Raskin: That means right now, only OpenAI-

Tristan Harris: Can make GPT-4.

Aza Raskin: Can make a GPT-4.

Tristan Harris: It's posited to have cost about $100 million.

Aza Raskin: That's right. And so then the question is how many people by 2025 are going to be able to make their own GPT-4s? And the answer came back between 10 and 1,000. That's a big jump. And then the next question we asked was, what is the likelihood that GPT-4 would be able to run on a single laptop?

Tristan Harris: Right.

Aza Raskin: Really interesting question.

Tristan Harris: Because right now it only runs on OpenAI having to pay some big cloud provider lots and lots of money per month to run what's called inference, which allows that blinking cursor on chat.openai to run GPT-4. Only OpenAI has the model for GPT-4 and only they're running it right now.

Aza Raskin: That's right. And so all of the work they would put into aligning it, making it safe, that only works because it's running on their server, and hidden behind an API. If it's running on people's laptops, there are none of those controls guaranteed. What it came back with is that there's a 50% chance, according to these researchers, that GPT-4 will run on a single laptop by 2025 and a 90% chance that it'll run on a single laptop by 2026. Gives you a sense of how quickly things are moving. Then just one other thing for people to hold in their mind, and this comes from a group called Epic AI that does research into how quickly AI is moving. And they're asking, okay, how much more does $1 get you next year than this year in terms of compute?

Tristan Harris: When people think of there's GPT-4 and everybody knows eventually there's going to be a GPT-5 and a GPT-6, and those are going to be 10 times bigger each time the number goes up by one.

Aza Raskin: Yeah.

Tristan Harris: One of the questions we ask is if GPT-4 is going to be able... What makes it safe is that OpenAI can lock it up and try to make it say nice things and they kind of lock it, only they can run it. But when everybody can run it on their own laptop because the costs are lower, it takes less processing power to run it on your own laptop because the algorithms get more efficient, because it takes less data to train. What we care about is how much more efficient is it for smaller and smaller actors with less and less resources to be able to make something as powerful as GPT-4?

Aza Raskin: Mm-hmm.

Tristan Harris: And to do that we have to track basically how quickly are things moving to make the compute, which is how powerful your processor on the computer is, how efficient your algorithms are, and how much more money is being spent every year on training runs?

Aza Raskin: Yeah. If you think about how much more powerful each machine is, these things are getting on order 1.3 times more powerful every year. You then think about how much more efficient the algorithms are, and that's 2.5X. And then the final one is, and how much more money is being spent? And that's 3.1X. If you sort of multiply all these things together, what you get is that things that took $10 today are going to cost $1 next year. That means if you have the capacity to train a GPT-4 for $100 million, I think next year that's only $10 million.

Tristan Harris: Wow.

Aza Raskin: You can really see how it just more and more people can both train and run these systems. It's going to be increasingly difficult to contain, because every year the wave gets 10 times more powerful.

Tristan Harris: And cheaper.

Aza Raskin: And cheaper.

Tristan Harris: And more voluminous.

Aza Raskin: Yeah. That's the bad news. But luckily we have some good news too, right?

Tristan Harris: And I want to say it's not good news in the sense of we figured it out, it's all going to be fine. We're going to contain all of AI. That is what we need to do. We need to create some method of controlling this power being unbound from who's wise and responsible enough to use it. We do need to care about containment and what control structure can hold this coming wave of AI proliferation.

Aza Raskin: What momentum do we have towards where we're going?

Tristan Harris: Just to say, I remember you and I sitting here and talking back in February before we did The AI Dilemma about, gosh, we have to get a meeting to happen at the White House and President Biden needs to invite the CEOs of all the companies together to actually talk about norms and just setting commitments. It's almost like getting all the different labs that were building synthetic biology to get together and say, let's set norms so we don't accidentally create bio-weapons. How can we make sure we don't create that as an outcome?

Aza Raskin: Yeah.

Tristan Harris: And we used to say, how could we ever get that to happen?

Aza Raskin: And in fact, I remember being at the White House with you talking with someone there and just seeing the look on their face of like, oh my God, AI. This is yet another problem. We're dealing with the Ukraine war with Russia. There are so many problems. What do you want?

Tristan Harris: Yeah.

Aza Raskin: He was sympathetic, but he was like, that's not going to really happen.

Tristan Harris: And he was not Biden, just to be clear.

Aza Raskin: That is true. That person was not Biden.

Tristan Harris: And to say that I think it was in May when Vice President Harris actually did have the CEOs of major AI companies sit down at a table and it looked like it was, you could say it's just a press release and a photo opportunity, but a few months later, the White House did announce voluntary commitments from the lab leaders. This is basically the CEOs of Anthropic, Google, Facebook, committing to a bunch of agreements about how they're going to have safer security practices, more investments in alignment and safety research. These kinds of basic things. Now that's not binding with law, but going from a world where this wasn't on the agenda, the public wasn't talking about it, to a world where I think it's, what, 80% of the public is concerned or alarmed about AI.

Aza Raskin: Yeah. It's eight to one people would prefer we move slower, not faster with AI. Yeah.

Tristan Harris: I remember when we worked on The AI Dilemma, it was before the six-month pause letter.

Aza Raskin: Mm-hmm.

Tristan Harris: And we started working with the Future of Life Institute, which actually did do that six month pause letter, and that made international headlines. The fact that eight to one Americans would prefer that we move slower, not faster with AI is kind of in the same ethos and vein of moving the public sentiment, right?

Aza Raskin: Yeah.

Tristan Harris: And we should celebrate that.

Aza Raskin: Something else that happened is you met someone. Who did you meet?

Tristan Harris: When President Biden came to San Francisco in June to meet with civil society leaders on AI, I met with President Biden to talk about a lot of the things that we brought up in The AI Dilemma presentation. And what's powerful about that is actually Gavin Newsom's, Governor Newsom's team was in the room. His team we know and Biden's National Security Council and Office of Science and Technology Policy and the President himself. There's a lot of different groups that are basically activated on these issues.

Aza Raskin: What was something that really bothered you about the meeting and also something that really made you hopeful in that meeting?

Tristan Harris: Yeah. One thing I can say is that the President and Governor Newsom and many of the politicians that we've talked to are all very worried about truth, trust, and democracy. The United States is the only country that is based on an idea basically, right? It's not based on a specific people, it's a melting pot of lots of people. And so a country that's backed by an idea is far more vulnerable to that idea being shaped and moved by information. And I thought that was actually a really interesting thing that President Biden spoke to. It's much more easy to manipulate or make people feel bad about an idea when you're sort of able to distort it with synthetic media. Or say, make a fake video of Biden saying, we're going to declare the draft when he didn't do that.

Aza Raskin: Right.

Tristan Harris: And I think that politicians are already feeling like there's such low trust in institutions partially due to the 10 years of the first contact with AI, which is social media, because what gets amplified, the thing that's the most cynical take on what any institution did. And having seen the cost of that, and then you pile on AI to this, I think people are really, really worried about democracy in the next election.

Aza Raskin: Mm-hmm.

Tristan Harris: I will say that when I introduced myself, President Biden heard the Center for Humane Technology and he briefly joked, "Is that an oxymoron?" And I think I pushed back that I actually believe that it's quite possible to make humane technology in reference to your father's work on the Macintosh.

Aza Raskin: And we got a call from someone. Do you want to tell that story?

Tristan Harris: Sure. Well, for listeners who might remember, we in The AI Dilemma talk, I think I opened the talk by saying it felt like in this moment it was March 9th, 2023, and we're talking about all the risks that are going to come from this. And I remember when I was telling the audience that we got calls from people inside AI companies telling us to make this presentation, that it felt like getting a call from J. Robert Oppenheimer who led the Manhattan Project. And imagine you have no idea what an atomic bomb is, and you get this call from a scientist who's telling you about this thing where the whole world's going to change. Literally, it's not just a weapon, it's going to change the world structure.

Aza Raskin: Mm-hmm.

Tristan Harris: And how do you take that seriously as someone who hasn't even oriented their mind to really feel through and think through the consequences of what this person's really telling you?

Aza Raskin: Mm-hmm.

Tristan Harris: And we referenced that as a metaphor in the talk, but then actually after The AI Dilemma went out, some little piece of good news is actually some family members of the Oppenheimer family reached out by email to us. I remember one of our donors who actually supports our work connected us. And the Oppenheimer family actually offered to host screenings of *Oppenheimer* with people from the technology companies, the AI companies. And they are very worried about what AI is introducing to the world is very parallel to the creation of the atomic bomb. And a lot of people at the AI companies that we know here in San Francisco did go to see it. Famously, Sam Altman, who's the CEO of OpenAI said that he actually was disappointed in the film because he thought it was a missing opportunity to get people excited and inspired about physics, rather than to really tune into the gravity of the creation and the consequences.

Aza Raskin: Yeah. He then also said that he thought *The Social Network* did a really good job because it got a whole bunch of people to jump in and make new social networks and apps. And often, I find Sam Altman has a nuanced take. This seemed just like the very worst possible take.

Tristan Harris: Yeah. It was a bit disappointing because I know Sam, I've talked to Sam in the past about social media and he deeply endorsed our view on what caused the race to the bottom of the brainstem and this competition for attention. He knows the problem of social media, and here he was validating *The Social Network* as saying, *The Social Network* was good at getting people excited about building more tech in Silicon Valley. It's like, no, it didn't. You should be smarter than that. You actually went, I forgot. Aza went to the screening of *Oppenheimer* with the Oppenheimer family. How was that?

Aza Raskin: One, it's just sort of crazy to be sitting there with the grandchildren of Robert Oppenheimer. And when the film ended, and mind you, we saw it in IMAX. I don't know how many stories, it's an eight-story tall. It's a very immersive storytelling. And there were a whole bunch of AI people in that room. And when the lights came on, there was a very uncomfortable silence. Just sort of this palpability of everyone not knowing what to do. In fact, everyone stood up, and then everyone sat back down again, and then everyone sort of stood up and then there was milling around. It's very clear that there was something very visceral that happened.

Tristan Harris: I think in summary, just sort of say what these shifts are. It's like, people can look at this very bleak situation and it is incredibly bleak. And we just came from a three-day workshop where things look even more bleak. But you have to also point your attention to the things that are shifting.

Aza Raskin: Mm-hmm.

Tristan Harris: It was not the case that there was going to be a White House meeting. It was not the case that *Oppenheimer* was going to come out and have all these AI lab leaders sitting down with the Oppenheimer family. It was not the case that Snapchat... Actually, we talked about the fact that they had this, My AI that showed up in Aza's fake 13-year old account on Snapchat when he posed as a 13-year-old girl saying, I have a 41-year-old male boyfriend who wants to take me out of state to have sex for the first time. And it gave, I'll just say bad advice.

Aza Raskin: Music and candles was the advice it gave.

Tristan Harris: Yeah. It was recommending to have music and candles for your first time to make it romantic.

Aza Raskin: Yeah.

Tristan Harris: Great advice. This actually turned into a *Washington Post* article that ended up going viral and senators in Congress have been resharing that article. We got contacted by several of them. And that was because you made that demo, you signed up and said, let me show you that this model is not safe.

Aza Raskin: Mm-hmm. Yeah.

Tristan Harris: And I want people to know these stories because it shows that if we can point to the harms, if we can point to the risks, if we can create a new social norm that it's not okay to just ship these new untested, large language model AI, golum AI's into your 13-year old's pocket, don't do that.

Aza Raskin: Right.

Tristan Harris: And if you say, don't do that, and you make it clear, you can actually shift the direction of history. And that little example is one taste of that.

Aza Raskin: To the listeners I think may feel hopeless too. Often we get that feeling.

Tristan Harris: Yeah.

Aza Raskin: I get that feeling, just to be really direct and honest about it. But just imagine if there were 10 times more people doing similar kind of defense work, and then imagine after that there are 100 times more, and then 1,000 times more. It can have a real impact.

Tristan Harris: Yep.

Aza Raskin: There's one other good piece of news I think we should share, and that is Senator Majority leader, Chuck Schumer, has been organizing something. He's called the AI Insight Forums. And this is actually really interesting because they're trying to do something new that Congress has never done before. Normally when Congress is trying to learn about some new technology and the harms it might create, what do they do? They ask people, a couple experts to come in and testify. Every senator or Congress person sort of gets five minutes to ask questions.

Tristan Harris: Which they're mostly doing to create a social media clip.

Aza Raskin: That's right. It's about making a thing that goes on CNBC or Fox News. It's really performative. It's not really about learning. And so what they're doing now is they're saying, all right, let's not do that. Instead, we are going to invite a set of experts to come deliberate. The opening plenary, they're thinking of having roughly 30 people or so, 30 experts, and then Congress sits around the edges, 100 members of Congress and Senate, and listens to this conversation about what we should do. It really is about learning in a profoundly new way, and I think that's really exciting.

Tristan Harris: Yeah. And we'll be participating.

Aza Raskin: That's true. And we'll be participating.

Tristan Harris: We were invited to join for the opening Insight Forum, which will be on September 13th.

Aza Raskin: That's right. I'm excited to see how that goes, but I really want to commend Schumer and also Congress and the Senate for doing something new, realizing that the rate of speed of this technology is so quick that they have to learn in a new way and doing some innovation.

Tristan Harris: Just to give people another taste of Aza and my work in this space, we also sit down with people who are at the companies.

Aza Raskin: Yeah.

Tristan Harris: And we found that even at pretty high levels of the company, people are very concerned. There's actually this point in the conversation where sometimes people will just sort of say, well, if I really could, I would just shut it all down and not have people build these advanced frontier AI systems.

Aza Raskin: Important to note, when people say shut it all down, what the AI community means is shut down the frontier. The largest models, the next -

Tristan Harris: Like the GPT-4's, five, six's, the biggest stuff we've ever made. It kind of reminds me of saying, let's not build the hydrogen bomb.

Aza Raskin: Right.

Tristan Harris: We already have nuclear bombs, let's not build the hydrogen bomb. We should explain this concept of when people say shut it all down, but they don't mean to shut down all AI and don't build it at all or don't have open source. What they mean is these really dangerous systems, these future dangerous systems that might be 10 times, 100 times smarter than humans, maybe they can do science on their own and they can do their own science experiments with robots and chemistry and they can start synthesizing things that we've never even thought of. That's not that far away. That's not too many steps ahead of the kinds of systems that we already have right now.

Aza Raskin: Dario Amodei, who's the CEO of Anthropic, one of the major players in a recent interview, he said that human level artificial general intelligence is two years away. And when we've had conversations with people at OpenAI, they say super intelligence, that is better than human output across most economic activity, that is four years away. Just give a sense of what the people inside think in terms of timelines.

Tristan Harris: Yeah. When we say shut it all down, what would we actually do? What would be the button that we're pushing and what would that cause? In this world, you wouldn't say get rid of GPT-4, the existing systems that we have. You'd say, okay, let's imagine they're training GPT-5 in the lab, and within the labs, they have this set of things called evaluations or evals. If you're running these evaluations, what you want to test for are dangerous capabilities. Does it know how to deceive a human? Can it successfully deceive a human? Does it know how to take its own code and maybe make it better? Does it know how to exfiltrate its own code? Can it steal its own code and get it to run on another Amazon web server?

Aza Raskin: Yeah.

Tristan Harris: Could it make a certain amount of money independent of human involvement? These are the kinds of tests that you, it's not all the dangerous ones, but these are the kinds of tests that start to say, okay, this model kind of has a lot of capabilities. It's kind of a really smart kid. And smart kid's been trained on the entire internet and everything humans have ever said, written or done, this is kind of dangerous. The alarm bells are going off.

Aza Raskin: Yeah.

Tristan Harris: We should probably hit stop. I imagine the metaphor in my mind for this is you're Homer Simpson in the nuclear plant.

Aza Raskin: Yeah.

Tristan Harris: The red alarms are flashing red. Sam Altman gets the call. The question is, what would the labs do in that environment? And the Homer Simpson, it's like you smash the glass and then you look, there's no red button.

Aza Raskin: Right.

Tristan Harris: No one knows what would actually happen in this event.

Aza Raskin: Yeah. Just cobwebs and little spider scurrying off.

Tristan Harris: Yeah. This is not really a good state of affairs. A simple thing that should happen in the next few months before the end of the year, and we've talked to people about this, is we should host pause workshops for basically pause.

Aza Raskin: Pause gaming.

Tristan Harris: How do we practice pausing?

Aza Raskin: Yeah.

Tristan Harris: And we show us a workshop that says, okay, say you're OpenAI, say you're Anthropic and you need to pause. Let's game that out. What do you tell your board? What do you tell your investors? What do you tell your employees? What do your employees do while you're pausing? What do you tell Nvidia in which you already spent a billion dollars on the next chip order to have all the next chips come and you took out a loan for that? Now you're pausing, you're not making money maybe during the pause. How does this all work? And I think that we can develop those plans, but we need to do that urgently. It's sort of like we're Wile E Coyote and we're rushing off the cliff and we're like, maybe we should build a plan for when we need to look down. It's like, let's build a plan now and let's also make sure we're really clear on how far we are off the cliff.

Aza Raskin: Yeah, it's just important to note that so listeners can track, when people talk about when AI folk specifically talk about shutting it all down, what they're referring to really are third contact harms.

Tristan Harris: Yep.

Aza Raskin: This is when AI starts to gain these capabilities where it gets better on its own and you get this runway explosion of intelligence. All of that doesn't solve the problems that we focused on in The AI Dilemma, the second contact harms. And really what we're saying is, well LLAMA2 is out, Falcon is out. We need the time before the next major set of capabilities comes out, to try to shore up our open societies or democracies from second contact harms. Which is, by the way, very hard.

Tristan Harris: Then people's minds start to spin and they say, okay, I'm overwhelmed by all this because let's say we could get the US labs to pause, but we just said that the United Arab Emirates is releasing Falcon, the next open source model, and they released that a few months ago now, and they're going to scale it another 10X. Are they involved in those conversations? And this is really the question. This is why in The AI Dilemma we've referenced sort of global nuclear arms control is the metaphor for managing proliferation of AI. Except instead of uranium, it's running on chips, on GPUs. Now what people need to know is that there's this very limited window in history where essentially two companies, NVIDIA and TSMC, make the chips that are used for training the most powerful AI systems in the world. Two companies. Could the US government say we need to start controlling and monitoring the flow of these major chips.

We start getting a handle on where are people training, not like the GPUs in your MacBook laptop right there, not those. People's personal computers are fine. This is not about government surveillance of that. This is about specifically saying, could we track these most advanced chips and where they're flowing in the world? And there's only so many places, so many countries, so many labs where people are using these chips to make these most dangerous systems. But we have a very tight window in which a couple of governments and a couple of companies really have sort of a choke point on the supply chain.

And we already saw the Biden administration did the export controls on chips, The CHIPS Act, in which they're starting to restrict the flow of chips from Nvidia and TSMC to China for the most advanced chips, specifically for military technologies and quantum and other things like that. We're kind of in the proto steps of this, but we really need... I mean with the workshop that we were just in recently, the conclusion was, how would you get something like a global monitoring system of chips in basically the next 12 to 18 months?

Aza Raskin: Yeah.

Tristan Harris: We need to do it incredibly quickly.

Aza Raskin: Yeah. And really I think this is a good time to transition into talking about the AI, we're calling it the end games workshop because we're trying to ask some of the smartest technical minds and some great policy minds, what do we need to do to get to a world we actually would want to live in, given the actual state of the world?

Tristan Harris: Just to kind of recap this for people, when we ran this workshop for three days with the top AI safety people that we could gather into a room to map out what are the possible best case scenarios, the non-catastrophic scenarios, and how do you get to those? No matter which of those there were, there's only so many of them. The point is that all of them rely on locking down chips.

Aza Raskin: Yeah.

Tristan Harris: I want listeners to think about that, I want governments to think about that, I want national security folks to think about that. Because there's really a very tight window in which, for example, China does not have its own domestic production of these advanced chips yet. The US also does not have the advanced production yet, for the advanced chips. Really there's this limited window in which something could actually happen.

Aza Raskin: No, I think we should talk a little bit about, because people might hear this and say, lock down all chips, all compute. Are you just going to take away my computer? I'm using my computer to just run all the things I want. No, no, no. What are we actually talking about when we say that?

Tristan Harris: It's nothing like that. It's just a lockdown. Like many, many chips that are used in one place of the most advanced chips for these advanced training runs. Literally OpenAI will spend probably a billion dollars training GPT-5.

Aza Raskin: Right.

Tristan Harris: They'll get a billion dollars of these chips and they're going to be spending months of just running them and churning them to create what will be GPT-5, which will be a more intelligent entity than humans have ever talked to living inside of a machine.

Aza Raskin: Yeah. And just to note on the timeliness aspect, remember the folks in the AI workshop believed 90% chance probability that GPT-4 would run on a single laptop by 2026.

Tristan Harris: Yeah.

Aza Raskin: There actually is, there are two lines going here, which is the massive training runs which we need to lock down. And then we do need to think very carefully about how do we do essentially on chip, on computer governance so that the most dangerous capabilities, as these algorithms become more efficient and computers become more powerful, don't also end up running on a personal laptop in a way that doesn't break personal privacy. There was one final three hour session at the AI workshop that actually Tristan, you ended up leading.

Tristan Harris: Yeah.

Aza Raskin: I thought it was very interesting because it was asking 30 people to think through step-by-step, reason step-by-step, what would need to happen between now and 2026 to end up with compute control, compute governance. And to do it in the form of headlines. Imagine you're opening up the newspaper and every week you're opening up the newspaper and you see a headline as we move towards a safer world.

Tristan Harris: Mm-hmm.

Aza Raskin: And so we did. We now have a step-by-step set of headlines for what would need to happen. And I just would love for you to talk a little bit about that experience, what you took away from it.

Tristan Harris: Well, what it's going to take is, I think people want to look for an easy answer here, right? They want to say, oh my God, this problem is so bad. Can't Congress just pass a law and then I'll feel good and I can go home and sleep well at night? I think there's an interesting effect of 30 experts sitting in a room for three hours mapping month-by-month between September 2023 and September 2025, how did we succeed in locking down compute governance for the world, that was training these extra advanced frontier systems? And it was very sobering. The felt sense in the room was quiet, simultaneously appreciating the level of detail that I don't think that plan has ever been mapped in that level of detail.

Aza Raskin: I remember you asked that, has anyone seen anything like this plan? Has anyone seen this step-by-step? And everyone said they had not.

Tristan Harris: Right. Here we are where this is a frontier issue of civilization, and it carries enormous risk. And we have some of the top experts in the world and we're saying that no one has ever even put this plan together at the frontier of that wave. It's like you're riding the edge of a surf of wave at the end of history. And you're asking yourself, what is a plan for surfing this wave? And a helpful shift that's made for me is instead of seeing humanity on the precipice of a cliff, seeing humanity surfing the edge of a wave. And I think about you in Costa Rica, Aza, surfing your surfboard, and I think that we need to collectively surf this wave as a species.

This is calling forth a rite of passage for humanity. This is not going to be some easy thing, but that doesn't mean to give up. Every day you and I are waking up and we are asking ourselves, where is the leverage to get that 12 to 24 month plan done?

Aza Raskin: Yeah.

Tristan Harris: How can California enact stuff with insurance that can make stuff happen? How can employees sign a secret contract that says, hey, if the companies were to get all these red alarm bells ringing and we didn't hit pause, we would quit. And we can sign a contract saying that we won't reveal our identity, but we will all simultaneously quit if we don't pause. How can the national security and executive orders of the Biden administration take this seriously and make some aggressive things happening with compute? How can Nvidia and TSMC recognize these challenges and even though they have trillions of dollars of market cap on building the next version of these chips saying, how can we get this right and do it safely? How can we create a culture of safety at a more human and wisdom level? And how the technologists who are building this all operate like more of the Oppenheimers who after having seen the bomb saying, I created death, the destroyer of worlds. How do we say, we are creating enormous risk, and we need to get this right?

Aza Raskin: What I would say about seeing that plan written out, and it's less a plan and more like a plausible path-

Tristan Harris: It's a plausible path.

Aza Raskin: Is to use the wave metaphor. It's as if before we all knew we were sitting at the top of this wave, but it's dark and I can't see where the bottom of the wave is. And so it just looks impossible. It's just in there is magical thinking and hoping that something will happen and maybe I should start surfing this way or this way. I just don't know. And here it's as if there's one line of light that I can see, oh, there is a plausible path from here to the bottom of the wave where I don't get walloped.

Tristan Harris: Yep.

Aza Raskin: Is that the right one? Probably not.

Tristan Harris: Right.

Aza Raskin: But the existence of ones means that there could be more who maybe tried this little thing to be like, oh, there are paths possible.

Tristan Harris: Right. And that's actually something I just want to encourage people to think in because what was an unlock for this group that we brought together was seeing a pathway, end to end, from here to there in which we can get there. Because people, I think, have a tendency to look at their small problem, which makes sense, by the way, we need to push on small problems. But I think we need to also see as we push on the small areas that we have leverage over, whether it's, if you're a culture creator, can you make TikTok videos about this? If you're a legislator, can you rally people up and get them to see The AI Dilemma? Make it required viewing for your staff. If you're a teacher, can you mail your congressman or woman, and host a screening of The AI Dilemma and The Social Dilemma? Sure, why not? Do both. And then send people's attention to say, AI really needs to shift.

Can we get public polling to start showing that the consensus that we need to slow this down, that we're moving too fast to get this right. Can we cool down some of the full on arms race dynamic with China so that we can take seriously that they don't want to go too fast and lose control either. That we both have a shared interest in not going off the cliff. I do think that if people have a shared pathway that they can see of how we could get there, I want to have as many people see that and operate from that and think of other pathways. There's no ego in the pathway that we happen to get out of this group of 30 people. I'd love to see 50 other groups do their own version of that exercise. How would you get compute governance to happen in the next two years. And get the best collective intelligence of people who know all the different disciplines and policy stake holding at play, and imagine what that would look like.

Aza Raskin: Yeah. I think just to end this episode, I want to ask you a question that we often get asked. I'll give my answer as well, which is, all right, so given all of this, are you optimistic or are you pessimistic? I sort of hate this question.

Tristan Harris: Yeah.

Aza Raskin: And my answer is, normally I'm neither optimistic nor pessimistic, but I make room for hope because to not do so is its own self-fulfilling prophecy. But you actually gave me a different answer to this and I loved it and I really wanted you to share it.

Tristan Harris: Yeah. The answer, are you optimistic or pessimistic? I say, I don't think about that question. I think about what would it take for this to go well?

Aza Raskin: Yeah.

Tristan Harris: And you point your attention at that ruthlessly and with discipline every day, what would it take for this to go well? And if everybody asks themselves that question, and if everybody has more maps that are provided by more people, because more people are thinking through what that map needs to be. And everyone just focuses on what it would take for this to go well, we have higher chances of getting there.

Aza Raskin: I think that's actually a really good place to end this episode. Thank you so much for coming to *Your Undivided Attention*. *Your Undivided Attention* is produced by the Center for Humane Technology, a nonprofit working to catalyze a humane future. Our senior producer is Julia Scott. Kirsten McMurray and Sara McCrea are our associate producers. Sasha Fegan is our managing editor. Mixing on this episode by Jeff Sudakin. 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.

Tristan Harris: Do you have questions for us? You can always drop us a voice note at humanetech.com/askus, and we just might answer them in an upcoming episode. A very special thanks to our generous supporters who make this entire podcast possible. And if you would like to join them, you can visit humanetech.com/donate. 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.