Tristan Harris: Hey everyone, this is Tristan.

Aza Raskin: And this is Aza.

Tristan Harris: *Your Undivided Attention* is about to have its second ever Ask Us Anything episode. Are there questions that you'd love to ask me or Aza about the show or more broadly about our work at the Center for Humane Technology and how we want to tackle these questions of AI?

Aza Raskin: Here's your opportunity. Go to humanetech.com/askus, or, and we love this option, record your question on your phone and send us a voice memo at undivided@humanetech.com. That's humane tech.com/askus or undivided@humanetech.com. We hope to hear from you and we are really looking forward to continuing this dialogue.

Tristan Harris: In this episode, we wanted to share with you some insights from the AI Insight Forum, which was held this past Wednesday, September 13th in Washington DC. This was a totally unique thing that has never happened in history so far as I know in the Congress, in the US, in our democracy, which is that Senator Chuck Schumer, Senator Rounds, Senator Young, Senator Heinrich all hosted this unique forum in which all the senators that came in were listening.

Aza Raskin: Normally when Congress needs to learn about something new, they hold a hearing and a hearing looks like a couple experts sitting, then the senators will have five minutes each. They'll ask questions, but they're not really asking questions to learn. They're asking questions to get a ten-second soundbite on Fox News or CNN. Democratic Majority Leader Chuck Schumer had I think a really important insight, which is that's not the way to learn. They're going to have to do something new. What he did along with Senators Round and Young and Heinrich was innovate. They had a set of experts and we'll talk about who those people were, sit and sort of have a structured dialogue where 50, a hundred senators, Congress folk just sat and watched and listened from 10:00 AM until 5:00 PM. Tristan, do you want to talk about what that felt like and what happened?

Tristan Harris: I almost want to tell a joke about Elon Musk, Satya Nadella, Sundar Pichai, head of Google, Mark Zuckerberg, Bill Gates, Jensen who runs NVIDIA all walk into a bar. Well, it's actually more like they all walked into a senate congress room. There Aza and I are, we're in the halls of the Senate. I remember we're walking, we're just talking to our chiefs of staff. We've got a coffee in our hands and we turn the corner and suddenly there's just this flash, flash, flash, flash of cameras and people shouting and saying, "What are you going to tell Mark Zuckerberg? How is AI being done safely?" People are yelling these questions at us.

I remember how surprised I felt turning that corner because it suddenly hit me what we were about to walk into. Then I turned my gaze and I see there's Elon Musk, there's Bill Gates, there's Satya Nadella, the CEO of Microsoft, there's Sundar, the CEO of Google, there's Mark Zuckerberg, there's Eric Schmidt, there's Sam Altman who runs OpenAI, and then there's all these leaders of major civil society organizations. This has never happened before. It felt like a movie. This is bizarre. To be honest, I felt quite nervous and anxious kind of walking into that room.

Aza Raskin: I think we both did.

Tristan Harris: Yeah, because it's one thing to be working on these issues and talking about it for many years, it's another thing to suddenly be directly across from the CEOs that are presiding over it and both having to face them and them having to face us. Not that it's oppositional, just that we have something very serious to work out here. I actually opened my senate remarks this way as kind of Elon Musk hinted in this way during his opening remarks too, that this feels unprecedented. It feels like history in a weird way. This to Senator Schumer and Young and Rounds and Heinrich's credit, was kind of actually making that historic meeting happen, which is the thing that we've wanted. Just like earlier when we started working on the AI dilemma, we wanted the White House to invite the CEOs. We said, "That's never going to happen," and then it happened.

This is another thing that's kind of out of a movie and it needed to be this because that's what's actually at stake. That's really what this represents is. To me, there's many ways to see this hearing. There's many ways to be disappointed about what more could have been. There's many ways to feel like it's just a photo opportunity. I view it, if I want to view it as optimistically as possible, it was a forum that treated this moment in history with the level of attention and platform that it deserves. A hundred senators are not sitting around asking questions of the CEOs. No, they sat down quietly in these chairs in front of 20 or so of all of us, the experts that were brought in. I sat next to Bill Gates on my right staring all the way across the room was Mark Zuckerberg and there's Sam Altman, and then we got into it. I'm curious, Aza, to what you felt like walking into the room.

Aza Raskin: Yeah, well, I mean honestly I felt a little intimidated, a little less by the people that were going to be there and more by the moment, knowing that this is the point in the movie where things have to turn and if the turn doesn't happen, that's sort of it. This is going to be a tragic movie. They're going to be around 10 insight forms. If you imagine the music for the movie, it hasn't gone to the hopeful music yet. This is sort of like the strings doing their back and forth building energy. It could be the turn. I think after a couple of the CEOs spoke, after Jensen of NVIDIA spoke, Sam Altman spoke, Elon spoke, a lot of that nervousness actually faded away for me. I realized nobody else was speaking about the set of incentives driving the race. If you know the race, the result, you'll get into that when I ask you in a second about your opening statement, Tristan.

The things that we talk about on this podcast really are not being represented in the most important rooms. I felt very emboldened that when we speak, everyone in the room's head turns and listens because it's representing something that's incredibly important for AI to go well. I've been saying the insight forum went better than I could have hoped, but not as far as we need. Your retort is of everything we need is 1% of what we need, but went 30% better than we hoped. I think that's roughly right. There was a moment when a Majority Leader Schumer asked after all the opening statements, "All right, raise your hand if you believe that the federal government is going to have to regulate for this to go well." Every single person raised their hand.

Tristan Harris: All the CEOs.

Aza Raskin: Every single CEO. That was a really important moment of consensus. Of course, if we put on our cynical hat, the way this is going to play out where they'll say "Yes, regulate us. Yes, we need regulation, but not that regulation." I think in terms of vibe in the room, it definitely was more open, more civil than I was expecting. When the CEOs get their opening statements, they were all canned statements. It honestly felt like many of them were written by their PR department. This was not from the heart, "Here, I'm showing up at this moment in history and I am grappling with the problem." A lot of them felt sort of canned. Now that wasn't everyone. Sam Altman and Elon Musk, surprisingly, and Jack Clark from Anthropic, there was a little more grappling I thought than say some of the others.

One of my big takeaways was, you don't often see senators sitting hour after hour after hour. Normally you see them when they're giving a hearing and they're grilling people and you see them in their power, and here you saw them just as human beings sitting with hunched shoulders trying to take in an insane amount of information.

Tristan, how did you lay it out when you were approached for this moment of what do you say to the titans of this era? What do you say to them and to Congress? What's the most important thing to be said now?

Tristan Harris: I mean, what was crazy is there we are, and on the other side of the tables, there's more than $6 trillion worth of tech that is advocating for the accelerating deployment of AI. The stakes really feel very, very high. Okay, so my opening statement, what I focused on was, why are we all in this room? We're in this room because we want this to go well. We all want a future that's going to go well. The problem is this belief that the future is uncertain, that we don't know which way AI will go. If we don't know which way AI will go, we don't want to regulate because what if we regulate too early and we lose out on the promises of AI?

The strong claim I made in my opening statement is that to a degree we can predict the future, which is a bold claim to make in front of that room. I said, "The reason that we know that we can predict the future," and this is borrowing from Charlie Munger who is Warren Buffett's business partner, "If you show me the incentive, I will show you the outcome. Show me the incentives at play and I will show you the outcome you're going to get." "In social media," I said, "that's exactly what we were able to do." I said, "I'm going to make a strong claim that I think we can predict the future because what is the incentive of the current AI companies that are building AI?" We all know that Facebook can connect people with cancer to cancer support groups and it can help long-lost loved ones find the romantic sweethearts from high school. Is that Facebook's incentive? There's a difference between the positive benefits that a technology can have. AI obviously can do material science engineering and help us solve climate change, but is that the incentive of those AI companies?

The answer is no. I'm not saying that the AI companies are not going to do those things. We all want them to do those things. That's why we're all here. We want that future. The point is to say what are the current incentives that if left unchecked, where is that pulling us towards? With AI companies, the actual race that's pulling them, is to scale and deploy these new intelligent capabilities to society as fast as possible without appropriate safety. GPT-4 can pass the MCATs and the bar exam. A few years ago, you couldn't take three seconds of someone's voice and clone them and now you can. GPT-2 couldn't give you accurate answers about how to make biological weapons, but current models can. GPT-3 couldn't find vulnerabilities in code to hack exploits and GPT-4 can do that. Everyone's trying to one up each other to scale and deploy more and more capabilities. That is the race that predicts what will happen as this is going to go on. As those harms accumulate, they will overwhelm the institutions that we have.

Aza Raskin: One of the most dramatic points in the hearing, Tristan, was actually with you and your old sparring partner, Mark Zuckerberg, and in particular it was about open source models Llama 2. You had a surprise ally of Bill Gates, and I would love for you to just walk me through what Mark's position was, what you said, what happened.

Tristan Harris: Well, one of the things I talked about in the solutions section of my opening statement is I highlighted that we are going to need to put certain restrictions on releasing open source AI models with dangerous capabilities. I used the example of Meta and it was awkward doing this because there's Mark Zuckerberg sitting across from me. I said publicly in front of everyone in the room, including a hundred senators, that Meta's Llama 2 model, which they claim was safe, if you ask it how to make a biological weapon, its safety controls will have it deny responding to that. I said to the room that we were able with a single person on our team with $800 to remove Llama 2's safety controls. Jeffrey on our team specifically created something called Bad Llama to ask it how do I make a biological weapon? It answered how to do that. I just remember how it felt in the room when I said it, which was that there was sort of this hush and this quiet and this kind of gasp.

Aza Raskin: You were saying, look, the way Facebook is using open source smuggles in a whole bunch because open source used to mean safer because there's more transparent and more eyes on it. That is no longer the case. You're now hiding behind open source. What open source means with AI and large language models, is it is less safe because once you put it out, no one knows what capabilities it has. It's now out forever and anyone can fine tune it to elicit new specific dangerous capabilities. You, Mark Zuckerberg, are endangering us by just rushing to release these models open source. That was sort of the implication of what you were saying. Zuckerberg then, he actually didn't get defensive exactly, but his defense was, "Hey, look, those biological weapons that now Bad Llama is telling you about, well actually you can just find those with a Google search." That's when Bill Gates, I saw him get physically animated.

Tristan Harris: I did too. He was sitting right next to me.

Aza Raskin: He turned his placard up to be called on. He immediately gets called on and he's like, "That is incorrect. You cannot just do a Google search to find that kind of information."

A really important question of course is like, "Okay, well then why would Facebook release anything open source?" It seems like maybe that's not in their business interest. Put it behind an API and charge for it. The point being this is a result of the race, right? Actually, Facebook has no longer released the biggest and most dangerous open source model that's now the United Arab Emirates that released Falcon 2, which has leapfrogged Meta. Meta is racing to leapfrog them again. The reason why Meta is doing this is that they are not competing as well on the largest frontier models. For them, they need to find a niche in the ecosystem. They need their area to be able to dominate. They've been pushing on open source so they can get developers on their side so they can have people work on their models so they get a whole bunch of mind shares so that their stock price goes up because they are a leader in one of the areas of AI. That's their incentive.

Tristan Harris: There's actually other angles here too, which is it's a race to get the best talent. The more you release these cool advanced models that show your company has the coolest, most advanced open source stuff, the more of the engineers who are advanced in machine learning and the PhDs they want to work at your company.

There's another reason as well, which is that when they release an open source model, people don't need to pay for GPT-4 because now maybe I can use the free open source model that Facebook built that's equivalent to GPT-3.5 and I can maybe run it in the future on my own laptop. There are good incentives for them to do this. The question was, those incentives run up against safety. We have a history of Facebook unilaterally deciding for the whole world what is safe. If I wanted to really twist the knife, I mean ask the people of Myanmar or Ethiopia, whether Facebook has had a good track record in setting the line for what is safe for the rest of the world. Just to be clear those are places where Facebook, by the, I think the United Nations, was basically saying that they helped enable a genocide.

Aza Raskin: Why you were bringing this all up is because this is the thing that Congress has to step in to regulate, to create rules of the road, to have a referee because otherwise the companies and the ecosystem as a whole are going to fall into the race to deploy.

Tristan Harris: I want to name and use the word referee. Elon Musk even was actually advocating for the need for regulation. He said, "Even though I'm connected to a whole bunch of people who want to delete the FDA and remove these things," he says, "I agree with the FAA 99.99% of the time and I'm glad that there's an FDA." He says, "I think we need a government regulator for AI." I think that was really important because Elon is followed by a lot of folks who are more in the libertarian side of the world who are right to be very skeptical of government regulation. Even him saying, "We need some kind of referee, we need rules of the road and limits on open source," is something that I think people agree on. In fact, to Mark's credit, who is actually quite respectful in that dynamic, by the way, I think people wanted to make it as dramatic back and forth between he and I on retrospect, but he actually said, "I think we agree, Tristan, that there needs to be future limits on what open source models that we release."

Aza Raskin: There are a number of actors and companies in the room, I'm thinking of Hugging Face, Palantir, Eric Schmidt that really harping on this idea of the race with China. They were actually using UAE'S release of Falcon 2 to say the US risks falling behind losing in open source. Other countries are leapfrogging us. They were using this to say, "Don't regulate us. Well, we need regulation but still don't really make it real. We need to go as quickly as possible." In my closing remarks, I think I got to use a reframe that Tristan, you and I have been using a whole bunch, which is that we cannot let our rivals define what the terms of the race are.

Tristan Harris: The US beat China to deploying social media as fast as possible. What happened? We beat China to creating a mental health crisis for our youth. We beat China to creating polarization among our citizens. We beat China at enabling outrage engagement algorithms to drive an incoherent unraveling of shared reality.

Aza Raskin: This actually got a whole bunch of the senators, both Republican and Democratic, to nod along that we do not want to beat China to AI in the same way. I think this is a critical reframe because for so long as, "We have to beat China," is the drumbeat, then we will be moving at a speed, to use Satya Nadella's term for how fast they were moving when they were releasing GPT-4, which is frantic. As long as we're moving at a frantic speed, then we will weaken America. What we all agree on, is that we need to strengthen democratic open societies with AI.

Tristan Harris: Again to applaud the format at this hearing, imagine if we went back to the industrial revolution and instead of just racing directly into the industrial revolution and just going through all the disruption, you actually consciously had a conversation about how do we want to do this? If we had that conversation about how do we want to do this, maybe we could have avoided a hundred years of child labor. That was actually kind a direction that in talking to Satya Nadella, the CEO of Microsoft, I think he was very pleased at a genuine human level to see that we were having a conscious conversation about how do we want this revolution to take place. Now at the same time of course, Satya is racing and self describing the pace that they were releasing AI at, using the word frantic.

Aza Raskin: I think one of the biggest things I learned is that there is still a fear among both politicians and the AI companies to really go there when they're talking about what we've been calling third contact harms. That is when AI becomes recursively self-improving, when it starts to automate science, when you get an intelligence explosion. There's a lot of sort of pussyfooting around it. They just sort of intimated it. Elon said, "We need to take civilizational risks seriously."

Sam Altman said actually something I thought was one of the most insightful things of the forum, which was to point out how bad our intuition is about where things will be in the future. He said, "Imagine rewinding the clock to 2020 and you were asked to give a prediction of where AI would be in 2023. Would you have gotten it right? Would you have said AI would be able to take a sketch of a website on a napkin and turn it into a fully working webpage? Do you think that AI would be able to solve the MCATs? Do you think that AI would be able to draw photorealistic images that you can't tell whether they're real or not?" The answer is no. None of us had that kind of intuition.

Then he asked, "Okay, now sitting in 2023, if you project your mind forward to 2026 or 2029, do you think your intuitions of how far AI will be are right?" Let that sink into your nervous system for a little bit? Because the answer is again, no, that we are almost certainly underestimating the kind of progress that will happen when you're on a double exponential.

One of the frames I shared, and this is originally due to Ajeya Cotra from Open Philanthropy, is that it's like 24th century technology crashing down on the 21st century. Just imagine if 21st century technology came crashing down on the 16th century. Suddenly imagine the king is sitting around with their advisors and suddenly they have to deal with cell phones and radio and television and the internet all at the same time. Do you think that their kingdom would've held? Do you think that governance would've worked? The answer is obviously not. Why should we think that our current form of 21st century governance, our democracies will be able to hold? Unless we do something unprecedented, then they won't. I think that line, that frame of 24th century technology crashing on the 21st century, that certainly got picked up by a number of different senators.

Tristan Harris: Another thing that stuck out for me was Sam Altman, Eric Schmidt, and Elon and Jack Clark from Anthropic, they were really focused on what some people consider to be the sci-fi risks, but of the incredible dangers of how fast this stuff scales and where we're going. I think Eric Schmidt sort of famously said and it really caught the room, he's a PhD in computer science and ran Google and he doesn't know how the latest AI systems are working. That's because again, these systems have emergent capabilities where the engineers themselves can't predict it. I think that was really helpful for many of the senators to hear because they think of Eric Schmidt as the brilliant PhD who was CEO of Google for so many years, and if he doesn't understand how it works, that says a lot because the field is going so fast.

Those of you remember in our AI dilemma presentation, that's one of the reasons that we're so concerned, is because can you effectively govern something when it is moving at a faster rate than you are currently able to apprehend? It's like every time you try to turn the steering wheel for the car you're trying to manage, it's moving at a faster rate than your eyes are even currently picking up. Do you think that where you nudge the steering wheel is going to be accurate if it's moving at a faster rate than where your eyes are currently appraising of reality? That's one of the real conundrums with AI. Just to link this back for listeners in our work more than a year ago we talked about in this podcast, the complexity gap, that the issue here is that the complexity and speed and power of technology is scaling way faster than the level of complexity of our governance. That's the meta issue, no pun intended to Meta, that we have to solve.

Aza Raskin: That's actually one of the questions the senator asked. I think he actually said that almost directly, that if Eric Schmidt doesn't even understand this, how can we possibly regulate it? It's a great question. We've already talked about how if you know the race, you know the result, you don't have to understand all of the internals to understand how it's going to impact the world. We actually just had a meeting with folks at the White House where we laid out everything we've been learning about what are possible and immediate and long-term ways of binding AI and making it go at a pace that lets us get it right.

Tristan Harris: There was this interesting thing that I'll say, which is so much of the head nodding in the room to the comments that we made was based on, we've already seen this movie before. We saw it with AI and social media. It was interesting that after lunch in the second part, a lot of the people came back from the company side and they were kind of pushing back on the fact that social media had been this big problem. I think they saw that it was actually getting a lot of head nods from the room that we got that wrong, including for say, liability.

The writer Upton Sinclair said, "You can't get someone to question something that their salary depends on them not seeing." The sort of quote that Aza and I were kind of batting around the phrase was, people who've been Sinclaired, where their beliefs, their epistemology are basically just a predictive of the incentives that they operate with. How many people in a room when we're actually trying to govern a technology, are doing the thinking that is independent of their incentives? Politicians have their incentives, and CEOs have their incentives. If you think about what would actually entail good governance, like answers to how we govern a technology that are based on the truth value of whether open source is in fact safe, should be based on a clean epistemology, a clean sense of knowing what is true, that is unencumbered by incentives, either from the political side, politicians who have to get reelected or stick with their tribe and decoupled from the CEO incentive side. What we radically need in rooms full of governance is clean thinking.

Aza Raskin: I just wanted to thank everyone of you who's out there listening to *Your Undivided Attention*. Thank you so much and we will see you or rather hear you next time.

Tristan Harris: *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. A very special thanks to our generous supporters who make this entire podcast possible. 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. If you made it all the way here, let me give one more thank you to you for giving us your undivided attention.