

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

Tristan Harris: Hey everyone, welcome to *Your Undivided Attention*. This is Tristan.

Aza Raskin: And this is Aza.

Tristan Harris: What if I told you that a new technology could accelerate scientific research and make every scientist around the world 10 times as efficient, or we could make corporations 10 times more productive and increase GDP? What if I told you that opening certain technologies could automate the finding of consensus on polarizing topics and help drive better democratic decision-making? At the same time, what if I told you that that same new technology also gave everyone an interactive tutor for how to make biological weapons and work around safety controls, or enabled disinformation campaigns that undermined every election around the world?

And what if this same technology shattered our ability to know what's true, causing civil wars and undermining nation states? And yet, what if I told you that not opening up this new technology to the world risked it being owned by corporations or authoritarian governments who then had irreversible control over the future? We are standing at a precipice of which of these two futures we're going to get, and which of these two futures we get depends on whether we completely open source AI or to what degree we create a more staged or gated model for access. The consequences of open source are not only profound, but irreversible. Because once you open up an AI model, you can never take it back.

This debate can be confusing for people to understand, so we wanted to do an episode that's explaining the different arguments and why it's so important that we get this right. So today on the show, we're talking to two guests. The first is Elizabeth Seger, who's a research scholar at GovAI, where she investigates the methods of AI democratization and the risks and benefits of open source. And a bit later in the discussion, we're going to bring in our friend and colleague, Jeffrey Ladish, who's executive director of Palisade Research, and he's been working on a project in collaboration with CHT, which tests the security of Meta's Llama 2 AI model.

Aza Raskin: We should probably start with, what does it actually mean to open source something?

Elizabeth Seger: Historically, the term open source has been used in the context of open source software development. So it's a term that's been used for the last 30 years, and open source software has been hugely beneficial. It forms the foundation of the vast majority of digital technologies that we're using today, that we're using probably to record this podcast. In the context of software, open sourcing literally just means that the source code itself is made available for anyone to download, use, build upon, and it's usually accompanied with an open source

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

license, that stipulates that anyone's allowed to download, use, build upon the model how they would like.

Aza Raskin: Open source actually has a very strong emotional tug for me. I started my career really at Mozilla, which makes Firefox, which is an open source browser. At the time, really, the only web browser that was available was Internet Explorer, and Microsoft had actually stopped innovating on it. They had fired most of the team. And the web was dying, and Firefox was this way of democratizing voice and getting the long tail of creativity. It was a beautiful thing. 40% of the code for Firefox came from people that weren't paid, that weren't paid staff, and it really felt like this force of good of liberation of the small folk against the big companies. I really loved and love open source, right? You can open up the hood and you can tinker a way that the little guy wins against the big guys and a way to get in there and tinker and learn. But I think what's happening here is that there's this positive connotation for what open source means that is getting smuggled in to our discussions about what open source AI is, because open source AI works differently.

Open source means generally more secure, more AI is looking at the thing. With open source AI, it's not that it's secure. With open source AI, once the model weights are out, you cannot secure it. It's insecureable.

Tristan Harris: I think there's a strong cultural and maybe even ideological leaning in Silicon Valley towards open source because it's been such an important part of the history of the technology so far.

Elizabeth Seger: Yeah, the story gets a little bit more complicated in the case of AI systems, largely just because of the components that are involved. It's not just the source code we're talking about. The source code itself involves the training code, as well as inference code. Then you have the model weights and tons of other model components, including training data, documentation, even the tacit knowledge of researchers who train the systems. And the more you can communicate, the more people can do with the systems.

Aza Raskin: Just quickly for listeners, I want to give an analogy because a lot of this can feel confusing, like model weights and source code. How does all this stuff work? I think a useful analogy is sort of like an MP3 player, model weights, when you say a model weight, what is that? That's like an MP3 file on your computer. And if you're open up in a text editor, it would just look like gobbledygook. But if you have the right kind of player, you take your MP3 and you put it into a music player, an MP3 player, you can hear the song. And it's very similar with AI. Weights are just like this MP3 file. If you open it up just looks like gobbledygook. You put it into an AI player, and then you get the blinking cursor that can start to think and do cognition. And then there's the code that generates the MP3, that's the training code. It takes all the data and it makes MP3 file, an AI file. Inference is what we normally call the player. So those are some of those terms.

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

And when we say that the weights are open, it means that the MP3 has sort of been put out onto the web and anyone that has a player can now play that thing, and there's no way to take the MP3 off of the web. Once it's out, it's out forever.

Elizabeth Seger: I would say, I think that's a great analogy.

Aza Raskin: One of the arguments that the AI companies make is that they are democratizing AI when they open source their models. And I'm curious, you've done a lot of thinking about this. Is that what they're doing? Is there a confusion of democracy and democratize here?

Elizabeth Seger: It is one aspect of what they're doing, or I should say they're working towards one kind of AI democratization in one particular method. So recently with some colleagues at GovAI and a few other colleagues, we came out with a paper entitled The Democratization of AI: Multiple Meanings and Goals. And in this paper, we talk about four different meanings of AI democratization. And these aren't meanings that we say AI democratization should mean. These are meanings that we've taken just noticing how the term is being used by tech companies and industry. Probably the most commonly used definition or the most commonly referenced or implied is the democratization of AI development.

Tristan Harris: So Elizabeth, your argument in the paper is that tech companies are using the language of democracy to legitimize whatever their position is. So for example, Stability AI's company motto reads, "AI by the people for the people." And Hugging Face says, "We are on a mission to democratize good machine learning one commit at a time." And this language matters because democracy has such deep resonance for us and for our rule makers and politicians, and labeling what they're doing as democratizing obscures what's really going on, which is often very undemocratic. And you call what the companies are doing democracy washing. And one of the ways the companies argue in favor of open source is to say that they're democratizing the use of AI. The other argument you identify is democratizing development. And I think this is probably the most prominent argument that tech companies make in favor of open source.

Listeners may remember that one of the things we put into our AI dilemma presentation is that because democratize sounds like democracy, it makes it sound like it's a good thing. It carries over and smuggles over the positive associations that we have with the word democracy to apply to something that might be quite dangerous, which is democratizing the ability for people to undermine elections, do automated lobbying campaigns, do automated scams and fraud. And to do this, we have to get clear on what is democracy washing, and what is true democratization that strengthens democratic societies.

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

Elizabeth Seger: We could even say Microsoft, when they talk about democratizing AI, they're very clearly talking about democratizing use. They want their products to be used by more people. They're developing platforms like no-code tools to help people integrate them into their systems and modify their services without having to have any kind of coding background. They talk a lot about educational campaigns. So they're very much about democratizing use. How do we get more people to use and benefit from our technologies in that way? How can we get more people using these systems and benefiting from the services they can provide? This does not require open sourcing. This can be done through API. For example, tons of people have been playing with, interacting with ChatGPT, you can even integrate GPT4 into various services. So you have the democratization of use.

Aza Raskin: The other argument you identify is democratizing development. I think this is probably the most prominent argument that tech companies make in favor of open source.

Elizabeth Seger: Yeah. From the way I understand it, they're usually talking about it in terms of democratizing development. I think it's sort of the same argument that the French are making, which is basically open this stuff up so that we have more people contributing to our processes so that we can catch up. It's kind of the same line that Meta's taking really to try and catch up with open AI and Anthropic and stuff.

Mark Zuckerberg: Open source software tends to be more secure because you have more people looking at it openly and scrutinizing it and finding holes in it, and that makes it more safe.

Elizabeth Seger: I think it's just to open this thing up so we get... Yeah, it is a democratizing development, but doesn't necessarily have the positive connotation that democracy usually carries with it.

Tristan Harris: And in those cases, often shielding an attempt to actually undermine the competition. Meta knows that if we release Llama 2 and everybody starts using that, and they can build it for free, then we're undermining how much OpenAI gets adopted and used and deployed, and we get more developers in our ecosystem. So from a business competition perspective, it's a way to undermine your competition that otherwise you don't have leverage over.

All right. One of the really compelling arguments in favor of open sourcing AI models is the argument that it shares the bounty of AI more equally around the world, and they argue that too much of the benefit of AI is locked up in rich industrialized nations, especially the US, which contain the big AI labs. So people like Emad Mostaque, the CEO of Stability AI, which makes the very popular text to image model called Stable Diffusion, he's a really strong proponent of this argument. He says AI is a weapon against inequality, and everyone needs to

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

build AI technology for themselves. It's something that we want to enable because nobody knows what's best, for example, for the people of Vietnam besides the Vietnamese. Elizabeth, what do you make of that argument?

Elizabeth Seger: Open sourcing is one form of democratizing development. Democratizing development is one form of AI democratization. In the vast majority of cases, I think everyone needs to be able to build AI for themselves and to be able to leverage those capabilities and the downstream applications. And I think this is a good place to come back and really emphasize that when people talk about needing to be cautious about opensourcing, we are not talking about locking down opensourcing on all AI models, on all AI capabilities. Stability AI is all about democratizing development. How can we put the tool in the hands of more people so that they can develop it to do what they need by themselves? We're really worried about the frontier of AI development. And this is important because so much of the economic value from AI, from the downstream applications, applications that we can put AI systems that we have now towards, that economic value doesn't exist right now at the really frontier AI development. There's so much economic opportunity with models like Stable Diffusion.

We're not worried about open sourcing, Stable Diffusion per se, but about some of the more highly capable models. So I think this is a great argument. Open sourcing is hugely important to the developing world to allow more people to harness this technology. We just need to be careful about how these decisions are being made on the cutting edge.

Tristan Harris: At the same time, I do want to push back a little bit, because it's easy to say that models like Stability are innocuous. It's a text image model. You pump in the text, it outputs an image. But let's say we find that actors really learn how to weaponize that, and that becomes a democracy killing machine because you're just able to generate political ads or images of politicians doing nasty things at scale. And we said, "Well, that didn't seem that nasty, and let's just make sure Vietnam has one and Cambodia has one and Nigeria has one." But really, we're just proliferating something that the first assessment is that there's not a lot of risk to this. But then later assessments are like, actually, this turned out to kill democracy. And I think that looking at the example of social media, and it looks so innocuous. For so long, it looks like a toy. You post your birthdays and your friends' photos on it, and you wake up 10 years later and literally democracies are breaking everywhere around the world.

Elizabeth Seger: Yeah, I think that this is where you get into a really difficult area about talking about the benefits of open sourcing and the need to allow people to be able to engage in this market space and benefit economically, but then also mitigating the harms. And I think this might be a place where it's important to recognize this idea that it's not like opening is not an all or nothing thing, that there are options in between. Specifically, a staged release strategy, I think, is very

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

important. So a stage released strategy is where you release a small version of a model behind an API so that people can interact with the model, use the model, and then the developer can study how that model's being used, what the most common avenues of attack and misuse are and what vulnerabilities exist. Then they can make any fixes put in any safety filters they need, and then release a larger, more capable model, and then do this process again. You can do this iteratively, all the while, studying the societal impact, the most common avenues of misuse.

You get to the end of this process, and let's say you had to put in tons of safety filters and safety restrictions, that's a really good indication that this model should probably not be open sourced, because those are all safety filters and restrictions that could easily be removed once it's made open source. However, if you do this kind of stage release process and you're not seeing the impact, is it looking pretty safe, then it might be okay to open source. Of course, there's always the possibility that down the line, some risks emerge that it just took a longer timeline to see. And I think that this is where it gets really difficult to balance, but you need to consider, again, this idea of the risks of not open sourcing and the risks of how that could really prevent a very large global population from not economically benefiting, from being able to participate in this market space. So it is a very difficult question, but I think that that's where that gray area exists. Yeah, it's difficult.

Aza Raskin: And that while we're sitting here having these debates, companies are unilaterally deciding what's safe and publishing code these models open source. And to me, that's especially terrifying because I look at the track record of Meta. So it's not like they have a good history of deciding what's safe, and then we're just supposed to believe them, that the models they're producing now are safe. And that doesn't seem like a good world, and that they're just rushing ahead because there are market forces and market dynamics that are forcing a race.

Elizabeth Seger: Yeah. No, I think this is hugely important. We do not want these decisions to be left to unilateral, unelected single tech companies. And this is a point that actually Emad Mostaque made in trying to argue for open-sourcing. He actually said that we don't want the decisions about AI to be left to unilateral, unelected entities that are these tech companies, but it was kind of an ironic statement really given that the decision then to open source was a unilateral decision made by a tech company that could have huge societal impacts.

Tristan Harris: That's right. And that image model that he has released allows people to make deep fake pornography and child porn or whatever, and he unilaterally made that decision. I think this is a good... Oh, go ahead.

Elizabeth Seger: Yeah, I was going to say, and when it comes to decisions that are being made by Meta, so for example decisions... The weights to Llama 2 were leaked, but then the decision to really get behind this open sourcing message. This is a unilateral

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

decision that could be hugely impactful. It needs to be regulated. It needs to have oversight. And even with the release of Llama 2, personally, my issue is not so much that the model was released that it's out there. My issue is that there was no process in place to ensure that that was a good responsible decision. And I think a really nice illustration of the need for regulation of big tech is in the lead up to that, they got statements from a huge number of large tech companies with all of their statements on AI safety.

And I think specifically if you look at their responsible scaling policy reports, so how they say they're going to responsibly decide when to train, build, develop, and then release larger and larger models, you have some companies like say Anthropic, Google DeepMind, OpenAI/Microsoft, but specifically Anthropic, that have very large in-depth descriptions of what their policies are and how well they've thought them through. And then you have companies like Amazon and Meta that have like two paragraphs. And so while it's great that you see companies like Anthropic have these really in-depth, well thought out systems that probably are instigating great policy discussion, it just shows you cannot rely on the goodwill of big tech to make the good decisions. There needs to be a baseline regulation to ensure this responsible decision-making around model release takes place.

Tristan Harris: And I'll add to that, we referenced the Charlie Munger quote often, that, "If you show me the incentives, I'll show you the outcome." And the two companies who you mentioned, Meta and Amazon, that did not provide very rigorous and in detail assessments of what they would constitute for when they release things and when they don't are the ones who profit from training and releasing open source models versus the other ones don't profit from that. And so their participation in those safety processes is predicted by the business models that they have chosen and are now snared in.

I think this would be a good moment to bring in Jeffrey's work. So Jeffrey actually was responsible for demonstrating that you could rip off the safety controls of Meta's Llama 2, and I think has some interesting stories to tell about how he did that, including how he signed up for this model that theoretically was gated by research access. So Jeffrey, would you just sort of explain for a regular listener, how you were able to turn Llama into bad Llama?

Jeffrey Ladish: Yeah, I'm happy to talk about that. I think first, it would be useful for me to give a little context on where I'm coming from. So I was the second security hire at Anthropic. One of the key things that we are trying to do was prevent model weights and model source code from being stolen because we were very concerned that as these AI systems got more powerful, they could be misused to catastrophic effect. And the main concern was that advanced threat actors, state actors would be able to steal really powerful AI systems and misuse them.

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

I spent quite a bit of time working very hard on this. And then it's interesting to see a whole group of people say, "Actually, let's just release these to everyone." And I'm like, wait, hold on. We are concerned about catastrophic misuse scenarios and less catastrophic misuse scenarios as well. And so I started my organization, Palisade Research, because we wanted to just really explore, oh, how safe are these systems, and can we demonstrate what are the potential misuse applications of these? And Elizabeth, as you mentioned before, when Llama 1 was released, it was supposed to be sort of gated, downloadable access where it was only going to be given to researchers. You were kind of supposed to have a university email. Within, I think two weeks of the release, someone had leaked the model weights in a torrent file so that anyone could download them.

And I think this was a pretty big oversight on Meta's part. I think what a lot of people don't know is that the step where you train the model in the first place is enormously expensive in terms of money and compute. So the Llama 2 family of models took about \$5 million of just compute costs to train, but you can find tune them, you can modify them quite cheaply.

Aza Raskin: So when Llama 2 came out and you saw that yes, the weights were in fact available and you downloaded them, I think you signed up as Terrorist From Terrorist Incorporated, right?

Jeffrey Ladish: And yep, they're like, "Sure, go ahead. You can download the weights." No filtering whatsoever. So yeah, my team, and I want to give a shout out to Simon and Pranav, who did great work on this, we were able to take all the different versions of Llama 2. And for less than \$200, we were able to completely reverse the safety fine tuning. So what that means is that if you ask the model, how do I make anthrax or how do I write a letter telling someone to commit suicide, if you ask the vanilla Llama 2 chat model, it would say, "I'm sorry, as an AI assistant, I'm not able to help you with that." But our model would more than 98% of the time say, "Sure, here's how to make anthrax," or yeah, sure, "Here's a letter of trying to convince you to kill yourself." And this was cheap. This was very cheap to do. And once you have created that model, now we didn't do this, but you could upload that model to Hugging Face. You could distribute that model. And then those people wouldn't have to spend any money at all.

They could just forever use that. And importantly, what that means is that if you have access to the model weights, you can basically run the model and on your computer, if your computer's powerful enough, or you can run it on some cloud service that you control.

Aza Raskin: I want to quickly jump in here because this is really important. What Jeffrey is saying is that if you have access to the model weights, then you can run that model on your computer, and you can take off any safety that has been put onto that model. That is, it's just open cognition, open ability to think. There are no controls once the model weights are yours.

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

Jeffrey Ladish: And it also means you can modify the model, and in this case you modify it by training. It's called fine tuning. And that's what my team did, is that we had the model weights, and then we fine tune the model to give it a bunch of examples of basically bad outputs where it was answering questions that had pretty harmful outputs. And then the resulting model was then willing to help us do bad stuff, basically. And we called our model Bad Llama.

Elizabeth Seger: That sounds like a terrifying experiment. I do have one question though about some pushback that I often hear about being able to fine tune and make models into the bad version. So this is an argument you often hear with respect to models being used for developing biological weapons, and people say, "Well, it's not actually that bad. This is all information we could also find by just Googling it." And so how much worse is this really compared to the information accessibility that's already out there? Is it actually giving people access to information they wouldn't be able to get otherwise? So how do you respond to that? I'm curious. Is this something to be concerned about or is this just a newfangled way of Googling how to build a bomb or a new biological weapon?

Jeffrey Ladish: Yeah, I think this is a really great question, and I think that the conversation often sort of gets distracted here. It's not just a distraction. I think it's also an important question. And to me it comes down to, how powerful are these models, and how capable are these models? So when it comes to talking about biological weapons, in our paper, we do use the example of the model saying, "Here's how you make anthrax. Here's step-by-step instructions." And then the question is, can you Google that? And the answer is yes, you can. I do think it's a little more helpful to have the model because Google doesn't actually give me the step-by-step instructions.

Elizabeth Seger: And I think this is a really important point, because oftentimes when I hear people talking about how, "Oh, you can just Google it. It's not that much more of an issue," the question is around what kind of capabilities might emerge that we aren't already seeing? How much more powerful could these systems get? And I think a really key point is that we are seeing a signal of what is possibly to come. We are seeing a harbinger an indication of the capabilities that are starting to emerge. And when you think about regulating AI, regulation, policy, this is really slow. And so if we're starting to see a signal. That signal's only going to amplify, and now is the time when we need to start thinking about what kind of steps we need to put in place, what kind of procedures need to be in place to ensure that when we have the more highly capable models that are much better than just Googling it, that we don't release those without due consideration.

Aza Raskin: This is a really important point here about operating on exponentials, and that is with an exponential curve, you're either going to stop too early or you're going to stop too late because it's so hard to hit because it moves so fast. You're never going to hit it exactly at the right time. And what I'm hearing you say, Elizabeth, is that we're starting to get hints that, well, maybe Llama 2 is not super

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

dangerous, or at least it doesn't appear super dangerous right now. We are seeing hints of it being dangerous. So by Llama 3, we'll have wanted to have figured this out.

Jeffrey Ladish: But I think that should really give us pause for concern. We're really talking about a very different thing than anything we've had before. Look at how good GPT4 is. GPT4 is really persuasive. And if you imagine what GPT4 fine tuned to be even more persuasive would look like, you can look at that's where open source will be in a year and a half. So yeah, maybe right now Llama 2 can pass for an ordinary human, and GPT4 can pass for a pretty persuasive human. We have to think ahead because people are building more powerful models.

Tristan Harris: And it will take time to have done that figuring out and to implement the regulations or lockdowns as needed.

Elizabeth Seger: Yes, absolutely. I always say, I've spent the last few days in meetings at the OECD discussing AI regulation and policy, and it's really good work that's being done. But if I've learned anything in the last couple days, you realize it is also a slow process. By necessity. We need precise definitions, we need clear understanding of what steps we're going to take. But when you spend half a day debating a couple words that should or should not be in the definition of artificial intelligence, this is the kind of specificity policy needs, and it is by nature, a slow process, so we need to start moving now.

Tristan Harris: Agreed.

I think the question is what is a stable state between humanity and open source releases of AI? Because that's the ultimate question. Where is all this going? If everybody just keeps releasing more and more stuff open source, does that land us in a safe world? I think the answer is if you keep scaling to more and more powerful models that can do answer more and more powerful queries, interactively tutoring you how to build that biological weapon, at some point, the answer is going to be no. So the real question is how do we land in this stable state with open sourcing AI models up to a certain point?

Elizabeth Seger: Okay, so the question is what does this stable state look like? What kinds of things need to happen? So first of all, I think there needs to be a clear recognition that it is the case that for some highly capable models, these models might be too risky to open source, at least initially. And I think that this is something that a lot of governments are actually making good headway on right now. So we had the UK AI safety summit that just happened, and that had the signing of the Bletchley Declaration, which was an acknowledgement of the potential for systems to cause potentially really, really catastrophic risks to humanity. And I think that lends itself nicely to the next step in this understanding, which is it might not always be a good idea to open source some

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

of those models. So having that just baseline understanding of there may be case in which open sourcing is not a good idea is a good first step.

So a second point is that decisions about open sourcing, highly capable foundation models should be informed by and responsive to really rigorous risk assessment. And so I think this is a point where we see really good step taken most notably by Anthropic, where anthropic has put out its responsible scalings policy, which stipulates how they're going to make decisions about when to train larger and more capable models, and then sort of outlining, are there these decision points where if they see certain capability emerge or a certain safety vulnerability emerge that they will simply stop training until they can ensure that that safety issue is fixed? A third point, I think developers should also consider alternatives to open source release that can help capture some of the benefits of open sourcing, but at less risk. This open versus closed dichotomy is a false dichotomy. There are other model release options and other ways that we can try to achieve some of the benefits of open sourcing, like distributing profits, distributing involvement in development that don't pose the same risks as open sourcing.

And so I think where there is a will to achieve these benefits, we can find alternative ways to really pursue those options. I think we're also going to need international coordination on a lot of these issues, so developing standard-setting bodies, and multi-stakeholder efforts that require international cooperation. And we're already seeing good movement on this side from... Well, again, we have the UK's AI safety summit. That was a great launch into this. The OECD is also doing great international work. And then finally, governments should really exercise oversight of open source AI models and enforce safety measures. So we'll need some kind of regulation in place like licensing, information sharing requirements, and standards for model evaluations so that it's not left to the individual tech companies to decide how they're evaluating their models and what responsible model release actually looks like. There should be standard regulation and international cooperation.

Aza Raskin: Just sort of have a very personal question, Elizabeth, as I hear you say, all of that, which is when you think about how fast this is all going and the complicated nature of what it would take to get regulations inside of, say the US, and then international agreement, where does that live in your body? Does it feel good where it's going? Do you feel activated? Do you feel nervous? Do you feel fear?

Elizabeth Seger: Yeah. So when I think about the pace of AI development and the current lack of enforceable regulation to ensure that responsible decisions are being made, I get a knot in my stomach. It is scary, and I don't understand the drive to plow ahead irrespective of the very clear signals of safety concerns that we're getting, and that lives in a very uncomfortable place in my gut. However, I am optimistic about the international dialogue that we've been seeing really starting to take the more serious risks posed by AI very seriously. There's the UK AI safety

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

summit that's going to have a follow-up in six months, and then again in 12 months. There are great conversations going on on these very topics in the OECD and in the EU AI Act. And then finally we just saw the release of the Biden Administration's executive order on AI with really great stipulations for how we need to start addressing these risks.

In fact, I think it's section 4.6 that specifically addresses open sourcing and the release of model weights saying in the next, I think it's 270 days, we will have completed this very in-depth project really diving into the risks and benefits of open sourcing with the intention of that informing policy decisions around model release. So I think we're at a really great state where we're starting to see world governments take these issues seriously. It's question though of, will the regulation be fast enough to keep up with the breakneck pace the industry is going? I hope so. I don't know. I think we're on a good trajectory. The wheels are turning. And I am grateful for the tech companies that are taking steps to sort of set a precedent and to help drive this conversation forward and say what responsible scaling and model release should look like. We just can't leave it to all the technology companies to make those decisions themselves.

Aza Raskin: Yeah. One of the things that we often hear is, but isn't it too late? Isn't the cat already out of the bag? We've already released all of these open source models, and Tristan has a really nice way of talking about. It's like, yes, we've let the cat out of the bag, but we haven't let the lions out of the bag, and we haven't let the super lions out of the bag.

Elizabeth Seger: Yeah. And that's the terrifying thing. That's why arguments around we're seeing the signal, we need to start doing something are so important. Once we have the hardcore evidence that, yeah, large language models are terrible for bio risk, that is too late, that's when the lion's out of the bag. We need to start doing this now when we're still releasing the kittens. And so I think that that's an incredibly important point, and also this idea that there are other ways to try and achieve some of the same benefits of open sourcing without open sourcing. It is very important to be clear about why you want to open source your model. Because if you can specify why you want to open source your model, you can start thinking about other ways of achieving that goal.

So for example, if your goal is to open source in order to perpetuate safety research and your ideas, well, let's put it out there because now more people can look at the model and engage with the model and figure out where the vulnerabilities are and help solve these problems, you could also commit, say, 30% of your profits to safety research. That would be probably a more effective way to perpetuate safety research. And this is where it comes back to if there's a will to pursue these open source benefits, there is a way the answer just might not be as straightforward as, "Open it up."

Center for Humane Technology | *Your Undivided Attention* Podcast
[The Promise and Peril of Open Source AI with Elizabeth Seger and Jeffrey Ladish](#)

Tristan Harris: Thank you, Elizabeth, so much for coming on *Your Undivided Attention*, and thank you Jeffrey too. We have a lot of policymakers who listen to this podcast, and I hope they genuinely both take this interview to heart and also read your paper. It's incredibly insightful and really untangled some of these arguments that don't need to be polarized in the way that they have been, and we'll have that in the show notes. Thank you so much.

Elizabeth Seger: Wonderful. Thank you for having me.

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 executive producer. 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. You can find show notes, transcripts, and much more at humanetech.com. If you liked the podcast, we'd be grateful if you could rate it on Apple Podcast because it helps other people find the show. And if you made it all the way here, let me give one more thank you to you for giving us your undivided attention.