ROUGHLY-EDITED TEXT FILE

# CDT – LAUNCH OF NTIA'S PUBLIC CONSULTATION PROCESS ON WIDELY AVAILABLE AI FOUNDATION MODEL WEIGHTS

DECEMBER 13, 2023

REMOTE CART CAPTIONING PROVIDED BY:

ERIK OLSON, NCSP RDR CRC CRR



* * * * *

>> ALEX GIVENS: All right. Let's get started. Hi, everyone, I'm Alex Reeve Givens and I'm the CEO here at the Center for Democracy & Technology. We're thrilled to welcome so many folks here in person and also are grateful to all of you who are joining us over the live stream.

And we're particularly honored to partner with NTIA today in launching their public consultation process on widely available AI model foundation weights.

This is one of numerous work streams. We've counted over 160 launched by the Biden administration's AI executive order and the commerce department has 270 days to solicit input and submit a report and recommendations to the president on potential risks, benefits, and other implications of foundation models with widely available model weights and appropriate policy and regulatory approaches.

We're pleased that NTIA is starting this public consultation so swiftly and with a commitment to seeking broad public engagement. That engagement is important because the issues presented here are particularly complex.

CDT was one of the most vocal groups, urging the Biden administration to push forward with the AI Executive Order and urging the administration to act swiftly to address the ways in which AI is currently impacting consumers, workers, and people's civil rights and civil liberties.

But on questions like the approach to open source and AI, there are still hard issues to work through, and it's right for the administration to act carefully and deliberately.

We hope to surface a number of those complex issues today and we will with some

thoughtful moderating.  This includes exploring the important role of open ecosystems in driving competition; empowering user choice; and facilitating innovation in security, transparency, and testing.

It includes pushing ourselves to articulate the risks of open model weights with specificity, to assess whether limits on open release methods would really mitigate those risks and how.

Throughout, we'll push tours talk with nuance about different approaches to model release, recognizing there is no binary of open versus closed and there are steps responsible actors can take right now to reduce risks in their release approaches.

Importantly, this conversation isn't just taking place today.  NTIA will need thoughtful contributions throughout their consultation on these issues and additionally, policy makers and governments outside the U.S. are grappling with all of the hard questions we're going to talk through today.

I'll say that CDT is deeply involved in this work and we're eager to share information and ideas with others, too, as we build the community talking about these issues and engaging with thoughtful policy recommendations.

To start this conversation, I'm pleased to welcome Alan Davidson, assistant secretary of commerce for communication and the director of NTIA, who will give remarks before we transition immediately to our panel.  Thank you all again, and thank you, Alan.

>> ALAN DAVIDSON: Thank you, Alex.  Thanks for that introduction and for your leadership here at CDT and I just have to say as our society takes on these giant policy challenges that artificial intelligence brings, we are so lucky to have a more sophisticated set of civil society players on the field than we've ever had before, and I really believe that that level of excellence is personified by CDT and the work of your new AI Governance Lab, so congratulations on that and thank you again for hosting us.

We're here today because of the growing power of technology in our daily lives. Advances in AI and machine learning systems have captured the public's imagination and rightly so.  These new systems will impact nearly every corner of our economy and our lives.

The starting point for the Biden Administration has been that responsible AI innovation can and will bring enormous benefits to people, but we'll only realize the promise of AI if we also address the serious risks that it raises today.

And those include concerns about safety, security, privacy, discrimination, and bias; risks from disinformation, impacts on the labor market.

For these reasons and more, there's a strong sense of urgency across the

administration and among governments around the world to engage on these issues.

President Biden's AI Executive Order, the most significant government action to date on AI, brings the full capabilities of the U.S. government to bear, and the Commerce Department is playing a leading role in the administration's AI work.

The Department's leading efforts on privacy, security, safety, innovation, competition, equity, and IP-related -- intellectual property-related concerns.

Our colleagues at NIST are standing up a new AI Safety Institute, the Patent and Trademark Office is exploring copyright issues. Our trade administration is promoting trade at U.S. companies and we at NTIA are trying to do our part, we are doing our part.

We are on the verge of releasing a report around AI accountability. We're engaged on the policy efforts across all of these efforts of the department and one area that we'll be focusing on at the direction of the Executive Order is AI openness.

In particular, the benefits and risks posed by widely available model weights. This includes model weights that have been open source or otherwise, broadly distributed.

And note my use of the air quotes around "open source." We know what people commonly call open source in AI is very different than the way, what we mean in the context of open source software, for example.

In fact, I think part of our homework assignment, our collective homework assignment here, is bringing better precision to this conversation when we talk about things like open-source AI.

AI openness raises important questions around safety challenges and opportunities for competition and innovation. Like open source software years ago, the early conversations about open source AI have engendered fears about safety and misuse as a starting point. Just a few weeks ago, I was at a conference and heard a prominent venture capitalist argue in reference to open AI models that quote, you don't open source the Manhattan project.

So that's one, maybe an extreme statement of it, but there are clear concerns about the power of AI and the dangers of making the most advanced frontier models widely available without any restriction or safeguard against misuse.

On the other hand, we know and we've heard from people concerned about the impact on competition and innovation, if only a small set of players control access to the most important models. History has shown us that closed systems can undermine experimentation and prevent technological advances, undermine security.

And we know that technology's benefits can be distributed more widely when access to that technology is democratized.

This does not need to be an either/or debate, as Alex so rightly noted in her initial remarks. This is not a binary choice about how we release models. We can seek, we should seek policies that promote both safety and allow for broad access.

To do so, we're going to need your help. That is why I'm so pleased to join will you have today, as NTIA kicks off our public engagement in our review of AI openness.

This review will lead to policy recommendations that seek to maximize the value of open source AI tools while minimizing the harms, and we're eager to hear from you. What should we focus our work on? What are the important questions to answer? And I know we have an all-star panel today to help us start that conversation.

We're keenly interested in the pragmatism of any approach that we take. We want this work to be grounded in a technical, economic, and legal reality of how AI is developing and being deployed.

Today's consultation is the start, and it's just the start. We are anticipating more opportunities to gather public input coming shortly in the new year.

So in conclusion, I'll just say, safe and trustworthy AI innovation and use, it's an ambitious goal, right?

The challenges on all of these different projects before us, they're daunting, but I am encouraged by one thing, and that is really all of you. The conversations like the one we're having today, like others happening around the world, they're evidence of a level of engagement on technology that's quite rare in our recent history. Governments are stepping up at this moment, businesses and civil society groups and technologists are stepping up, and this is our moment.

The potential for technology to promote human progress has never been greater. Making the right decisions now will lead us to a world where technology works in service of a more open, free, equitable, and just society.

Together, I know we can build that better version and vision of our future. Thank you.

[Applause]

>> KEVIN LI: Good afternoon, everyone. Thank you, Alan for those remarks. I'm Kevin Li. I work in policy at National Telecommunications and Information Administration's Office of Policy Analysis and Development and I'm really delighted to introduce for all of you our all-star panel on AI.

And as I call your name, please come on up.

So first, we have Kevin Bankston, who is senior advisor on AI governance at CDT where he supports CDT's AI Governance Lab in its mission to advance robust solutions to AI risks. Kevin previously served in a wide range of roles across civil society

including at the American Civil Liberties Union and the Electronic Frontier Foundation. As a previous director of the CDT's free expression project and director of the OTI.

Next, we have Peter Cihon. Peter is senior policy manager at GitHub. He works on public policy to support software developer communities across the world with a focus on AI and cybersecurity. He's published research on the regulation, international governance, and labor implications of AI, and serves on the OECD network of experts on AI. Prior to GitHub, he served as researcher at the Center for Governance in AI.

Then part of the University of Oxford and consulted for the OECD AI Policy Observatory and Global Catastrophic Risk Institute.

Next we have Sayash Kapoor. Sayash is a computer science Ph.D. candidate at Princeton University's Center for Information Technology Policy. His research focuses on the societal impact of AI. He previously worked on AI in the industry and academia at Facebook, Columbia University, and EPFL Switzerland.

And last but not least we have Elizabeth Seger. 11th is a research scholar at Gov AI where she investigates foundation model release strategy and policy. 11th holds a Ph.D. in philosophy of science and technology at the University of Cambridge.

So I'm really delighted to have the chance to be up here with such a distinguished panel. And as assistant secretary Davidson mentioned, we're here today to launch the very beginning of our public consultation on our work under section 4.6 of the AI Executive Order, which asks us to assess the risks and benefits of dual-use foundation models with widely available model weights.

And, of course, that's a mouthful, and so I want to start with a level-setting question on the nomenclature and how it affects the substance of what we're working on, the EO comes with a definition of dual use foundation models which have to have tens of billions of parameters and meet some other thresholds.

But it doesn't actually define the term widely available model weights and as assistant secretary Davidson mentioned, we've been using the term open source in discussions about openness.

So how should we be thinking about the scope of widely available model weights relative to something like open source, which might carry other connotations? Elizabeth, do you want to get us started?

>> ELIZABETH SEGER: Sure. So for widely available model weights, as outlined in the Executive Order, it sounds like just publicly accessible model weights. I think that this and others will probably be able to speak to this, as well. It's only part of the definition of what we think about when we talk about open source AI systems.

And I think one thing with open model weights is a lot of the risks that we're concerned

about with open model weights, these risks come to bear when just the weights are open but oftentimes, the benefits of open sourcing are not fully achieved with only the weights available.

We need lots of other model components available. For example, training code, training data, model documentation, support with compute infrastructure, support tools, education.

And I think you were speaking to this earlier, having a conversation, I don't know if you want to expand on this theory.

>> SAYASH KAPOOR: Sure. So to me there are two key distinctions that come in. What I like to call open foundation models, which is closely aligned to what the Executive Order referenced to as widely available weights are models where the weights can be downloaded by a large section of the population.

So this includes models where a large number of users have been able to download these models and includes more permissively released licenses.

So the first distinction to me is the model weights are widely available.

The second distinction is if we move along the gradient of openness, what is traditionally referenced as open source would be models released alongside the code that's used to train them, the data that's used to train them and we have no use restrictions in the license.

So a key part of what constituted open source software has been permissively licensed software which does not have use restrictions on how users can use these models.

There are two key distinctions, open foundation models meaning the weights are available but open source models which I think leads to a much more expansive view of what types of assets are made available publicly. And as Elizabeth mentioned, a number of the key concerns have been at the weight step because in order to use these models for disinformation or biorisk and so on, the concerns arise as soon as you have access to the weights, but on the other hand, to answer scientific questions of interest, things about bias or discrimination, things about how secure these models are or whether we can even verify what data these models were trained on, we need access to other assets, things like the code that was used to train the model as well as the data set.

In my view when we're thinking about the risks, I think the key distinction is open versus closed. Whether we have the foundation model weights openly available. When we're thinking about the benefits we need to move along the gradient and think about whether the other assets required to train these models are also available, as well.

>> KEVIN BANKSTON: Great points to get us started. I certainly agree that it's

insufficient to have the weights alone to see a lot of the benefits. I think it's important to reference in these definitional conversations, and I think it was quite elegant of the Biden administration to use a new term to side step what is a controversial discussion in the community around what is exactly an open source AI system or model.

It's important to reference the OSI, the open source initiative, as the arbiter of the open source software definition which we could spend a lot of time diving into the minutia of. It's important to think of the four principles that undergird what has enabled everything.

The right to scrutinize the code, to run the code for any purpose, the right to modify and redistribute the code.

As useful guiding posts for open source software.

And then the key question that the initiative is grappling with is how does that translate in practice to open source AI models?

So the broader framing of widely available model weights or openly available AI components is a constructive start.

I would also say we have a lot of great things and we'll spend time touching on those points.

One aspect I do appreciate is moving towards precision on some of these risks. So the consultation as I understand it is focused on dual-use models of widely available model weights.

And that is to say the dual-use definition is to your point larger models, ten plus billion parameters that have particular concerns for weapons of mass destruction creation, for cyber offense, and for loss of control. And I think narrowing that conversation is a useful step because there's so much benefit to be seen and we can have a more precise conversation of the risks and dive into it as we go.

>> KEVIN LI: Well, that's a really helpful segue into -- we've been talking a little bit about the four pillars of open source and Sayash, you mentioned a little bit about how those might be, you know, one of the dimensions of open source or of model openness, but not necessarily the only one.

There are a lot of equities across the board in openness in AI, so are there potential -- what potential benefits or harms arising from openness are top of mind for folks right now?

>> PETER CIHON: I'll start with that one. And I'll start with benefits, because open source has been getting a bit of a bad name in this conversation and I want to remedy that a bit because speaking generally, I would say that the ready analogy to open source AI is open source software and we can look to the history of open source

software and its context.

We were in a very similar period in the '90s when there was a lot of fear, uncertainty and doubt around the proliferation of open source software. It was suggested that it was more insecure, that it was socialist, that it was giving away our innovations to our competitors, encryption source code was treated like a munition, although the courts ultimately did not like that.

But we today live in a world where the vast majority of websites are hosted on Linux boxes, open source Linux boxes with open source Apache web software. We access them through browsers including Mozilla's Firefox and Chrome which were what enabled meaningful competition against Microsoft's Internet Explorer. We access that on an open source Android phone which enabled multiple hardware companies to compete with Apple in the smart phone realm.

So we have a ready analogy to the benefits of openness in software, looking at that history. And although certainly, it has not been magic competition juice that solves all concentration problems, I would say that the world we have today with certainly a few walled gardens and a few dominant players, but an enormous long tail of innovation in web hosts and somewhere writers, that would certainly be preferable to the counterfactual which would have looked like AOL and Microsoft no offense to either of them being the main engines of the information economy and all of us having less access to information, less access to the economic benefits of these tools.

So that's how I think about the benefits of openness.

I'll add in terms of the risks that are most top of mind, certainly for CDT, our biggest focus and concern is discriminatory decision making by automated systems.

We think, although we're still working on this as part of Governance Lab's work, it seems that many if not most of the mitigations around specific decision-making systems would happen at the deployment layer, not at the foundation model layer.

And so I personally worry most about the problems that will come with open source AI that have also been intractable on an open internet, those being in particular illegal sexual content, fraud, and things like that, things that have been -- that have grown exponentially with the exponential growth of all of our other information technology, and I don't think we have clear answers on how to combat that while also balancing burdens on free expression and innovation.

>> ELIZABETH SEGER: So if I could jump in.

So first of all, I want to say, Kevin, that I agree with everything absolutely you said. I'm going to layer on top of it, though, with a few other concerns and ideas.

First, I'm going to start with the risks and the risks that have been forefront of my mind in

the work that I've been doing have been mostly risks around misuse and security concerns, so these are risks around foundational models and generative AI being used to build biological weapons, chemical weapons, produce malware and there's a really important point to be made here.

The marginal addition of risks that we're seeing right now from current AI systems is not that much more than being able to just Google this information, and I think that's a really strong argument.

However, I would add that a lot of people who are worried about open sourcing of models aren't necessarily worried about the models that we currently have but the capabilities that might emerge in the future.

While I completely agree that the marginal risk is currently small, I think prudence demands watching what happens on the frontier, looking out for the development of new capabilities and being careful about the decisions we might make around releasing weights as capabilities continue to develop, as we're able to integrate different systems with each other and as they become more user friendly.

Second, I want to talk about the software analogy. I think that this is a really good analogy in terms of illustrating the benefits of open source and showing that especially in this offense/defense balance, open sourcing has allowed defensive capabilities to grow, even though open sourcing does pose risks and vulnerabilities.

If you open source a system, we can find those vulnerabilities, we can patch them. This wide distributed community-based innovation is really important for safety and developing defensive capabilities.

I think, however, that we should at least consider the possibility that this analogy might not hold perfectly in the case of AI systems. One concern that I've been thinking about is that while in the case of traditional software, this offense/defense balance, obviously has tilted very strongly in the direction of defense that it might tilt more towards offense.

I'm not so sure how far, but more towards offense with the more highly capable systems that we develop.

The reason being that the issues, the vulnerabilities, the bugs, they're just as easy to find and exploit, but potentially much more difficult to patch, and to solve than the case of traditional software. AI safety is difficult, it takes a lot of money, a lot of expertise that we're pouring into it.

And it is lagging behind the rate of AI development. So that would be -- I'm not 100% sure how far towards offense we'll tilt, but I think we should be open to considering that that offense/defense balance might not work out in the same way the more highly capable the systems get.

>> KEVIN BANKSTON: I think that's definitely a possibility. I think it's also a possibility it will help on defense. For example, open source tools may be the way that we can develop better AI detection tools, for example.

And so I'm not arguing with you. I'm noting there's a dearth of evidence --

>> ELIZABETH SEGER: This is an area where we need more research because it's not as obvious in the software case.

>> PETER CIHON: And this is another reason that we're launching pilots, patching vulnerabilities at scale using large language models in critical code. We need more attention to the defensive use of large language models and AI. The government can lead on some of these pilots. Broadly we do need more research on this offense defense question, but generally on AI security and safety and that's a clear benefit of open source releases to date.

There's multiple repeated references if you spend too much time on Twitter as I do, from folks in the AI safety community acknowledging that research has really advanced because of the open release of Lama 2 to model oversight and model interpretability questions on the safety side and exploit on the security side from prompt injection that affect open and closed models alike.

So we should certainly be expanding those types of research and open releases to date have helped with that. It's important to recognize the contributions to reproducible science that open source and open science supports.

>> SAYASH KAPOOR: I think that's a great point and I want to quickly jump into some of the risks as well and to see what evidence we have today and where we stand on those.

So one of the main risks that has been talked about with respect to open foundation models is the risk of cybersecurity threats. In particular two key things that stand out, language models can be used to automatically find vulnerabilities in critical infrastructure and therefore, hack into systems more easily, that's one. And the second is language models will be used to create nowhere which can then be used to infiltrate other systems.

So let's dive into a little bit of the history of both of these because I think if you go to it from the perspective of this empirical history of how such transitions have played out in the last three decades or so, I think that could be very informative.

So on the first concern, on finding vulnerabilities using automated tools, this is not the first time that we have superhuman capabilities at vulnerability detection.

For the last three decades we have had these things called fuzzing tools. Fuzzing tools are used in cybersecurity to automatically find bugs. They sort of put in inputs to

existing software, try to tweak them and find vulnerabilities this way.

And they have been used to find tens of thousands of vulnerabilities in critical infrastructure. So why hasn't the world ended? Why haven't we come to a point where everything is hacked all the time? Well, the reason is that defenders have access to the same tools.

So fuzzing tools have been used predominantly by cyberdefense organizations and by people who are trying to protect critical infrastructure. One of the most prominent uses has been the initiative called OSS fuzz, which automatically finds issues in prominent open source libraries and alerts the developers and prompts them to fix these issues.

The same would hold true for language models, as well. We've already seen the use of language models on detecting these vulnerabilities. Google has launched a thing a security vulnerability detection model, which is being used to expand access to vulnerability detection tools.

And similarly for malware, as well, Google has developed a foundation model that can detect new forms of malware. You can upload a piece of software to check on a website called Virus Total, not an advertisement, but these types of tools have already started being used for defense.

And I think there is very little reason to suspect, at least for these types of issues, vulnerability detection would be any different for any automated tools in the future. And to peter's point, I think a lot of open models have enabled this type of research on how useful these tools can really be for cybersecurity detection and to also allow companies to deploy those models locally rather than sending their outputs to open AI.

Quickly, I'll jump into biorisk, as well.

So I think the question on biorisk is what is the marginal risk, as Elizabeth mentioned, it's true that today's models are not -- we don't have any evidence that today's models are much better than searching on Wikipedia. There were a couple of studies that came out of MIT a few months back and I think they panicked a little bit and said if you use Lama 2, you can create bioweapons.

If you search the internet, if you search Wikipedia, you can find the exact same answers. A report from the National Academies of Sciences in 2019 had the exact lines of information in fact, that was referenced in these reports.

And so unless we're also trying to set out to ban the Nationals Academies of Sciences from putting out reports and censoring Wikipedia, we're on a slippery slope there.

I think it's important to think about how we can take steps to protect ourselves against the risks of future models and there I think the EU also does a very good job of pointing

to the downstream attacks.

So it's not just the information on how to create a bio weapon is enough to cause a bioterrorism incident. The way it actually happens is people have to use that information to come up with these agents to create them in the lab, a process that often takes tens of years even for the most talented scientific research groups.

So the way to prevent these attacks is also in my view one of the best ways is downstream, in DNA synthesis facilities and incorporating AI into DNA synthesis units, which is a proposal that was raised I think earlier this month by the Federation of American Scientists and that's another way open source AI or open foundation models can help in defense by scanning for what types of synthesis instructions are being given and so on.

And tilt the balance in favor of defense rather than offense.

>> KEVIN LI: Thanks for the broad range of benefits and risks. A lot of our conversation so far is focused on thinking about the marginal risks over thinking critically and analytically about what the marginal risks.

What should the proper baseline comparison be? Is it the top closed weight models or, as you mentioned, Wikipedia or Google?

>> PETER CIHON: So there's two questions. The first baseline is what new risk is introduced by AI writ large? And I think it's important to understand that there are significant risks introduced into society with AI systems and Kevin, you were touching on some of those important speeches, CDT's focus of poorly understood systems that are used for making high stakes decisions and there's regulation needed to make sure those things are handled appropriately.

There's also a number of societal challenges from criminal misuse and we can think about misinformation, deep fakes, malicious activities that require a lot of expertise whether that's bioweapons, cyber, or others.

It's important to consider thus far that a lot of those misuses are raised, the number of folks who are able to do a lot of these things, it's supported by a user interface that lowers the barrier to entry to use the technology.

And so at the outset, AI introduces these risks, but these risks are introduced by both open and closed models alike.

There's reporting yesterday about concerns of deep fakes in the Bangladesh election ahead of January and the reporting points to specific commercial applications that are enabling people at low cost to develop these. It's not somebody in their basement using a locally run open source model.

But that's the second important question is what is that difference between the closed and the open?  The marginal baseline, it's worthy to think there are new risks that are introduced by openly available.

And I think you've already touched on some of them.

So if we're concerned about somebody taking model weights and being able to work in secret in their lab and build something that might not be surveilled actively as if it were through API, that in practice, you face a lot of challenges taking malicious steps beyond information retrieval that can be observed by law enforcement and addressed.

And a second marginal risk that I would say is the concern of undoing a lot of the protections that might be offered on either an API kind of context with filters, but in the locally run model, Lama 2, safety filters being undone.

It's important to call out here, you need some skill to undo these filters, these fine-tuning methods.

And while the cost of doing this is not prohibitively large, especially in contrast to creating some of these models, that is a barrier.

If someone were to maliciously fine-tune a model, and then seek to release that into the world to wreak havoc and help all the script kiddies do a lot of bad things, we live in a society that has institutions and platforms that have handled dual-use software and its distribution for decades.  GitHub has acceptable use policies that if somebody is to intentionally share a model or an AI system that is intended for harm, say for deep fake, generation for sexual purposes or terrorist content, that's in contravention of our policy and we take that content down.

You see this across the internet broadly.

And so the risk profile of fine-tuned models by malicious actors shared to the wide masses to wreak havoc is very narrow, and I think can be addressed by law enforcement.

We need to balance that against the benefits.

>> KEVIN LI: Elizabeth, you seemed like you wanted to jump in?

>> ELIZABETH SEGER: I wanted to jump back in on the biosecurity point.

And back to this idea that again what we're not necessarily concerned about primarily are the systems we have now, so there's this great report that came out by the Center for Long-term Resilience that outlined different uses of the way that foundation models are being used in biological design.

And it's not just about -- it's not just about protein folding; it shows all the way from initial

protein folding all the way through biological workflows down to experimental design and biological synthesis, and just throughout the entire process.

And showing how different systems exist at different stages of development for these different processes.

And the concern we have is that down the road when you can integrate these different systems that we might then have systems that really lower the barrier to entry. So people can create user interfaces, people can use these systems to carry out biological workflows all the way through.

But to be clear I think in the first instance, what we're going to see is that these systems, they just help people who are already very well-resourced and have lots of expertise.

So in the first instance, marginal risk, probably small, but I do want to keep our eye on the ball that in a lot of these conversations, we're not worried about the immediate marginal risk, but the things that may happen down the line and keeping an eye on that.

>> KEVIN BANKSTON: I understand that. I guess my question there and we've talked about this and we'll talk about it more, the risk that you're describing from the future systems aren't coming from the supermodels.

>> ELIZABETH SEGER: That's the one thing about biological risk. It's coming from the small models.

>> KEVIN BANKSTON: -- some of the other risks we've been mentioning. They don't actually require the largest models and so what is an appropriate line to draw is the line of 25 or 26 megaflops, depending on whether you're looking at the EO or EU. Is that even a rational line, is that worthwhile question when the risks that we seem are most worried about, which is information about how to create weaponry, or the kinds of harmful speech we've been talking about can be and will be generated by much smaller models where attempts at nonproliferation of those models will fail.

>> ELIZABETH SEGER: On the question of where do we draw that line, we need to think of two different cases.

So we have the case of models that we're worried about with biological risk which are smaller, more fine-tuned, specific models, and then we have larger language models that have the issues around mis and disinformation, loss of control risks, building highly persuasive AI systems.

So for systems that we're worried about with biorisk, I think they're having some -- concentrating more on the downstream, make sure people can't get their hands on the DNA synthesis tools and stuff like that, that's probably the best place to focus.

However, when we look at the risks and harms that we're worried about and oftentimes,

more speculative risks and harms by these larger models, I think this ten to the 25 flops that's been laid out in the EU AI Act, I'm very in favor of it.

I think we could probably all agree that having a compute threshold cut-off is a really bad proxy for deciding model capability, but the nice thing about it is it's a very clear line that can be written out very clearly in legislation so it gives a very nice clear starting point.

And then one thing I really like that the EU AI Act has done is it says this is our line, and we're just going to assume that if you're above ten to the 25 flops that this is a high systemic risk system, but if you can show us that it's not posing these kinds of risks or showing these capabilities that we're worried about, then maybe we'll drop you down to the tier below and if you're at that tier below, but your systems are showing sketchy capabilities that we're worried about, we're going to bump you up.

Having that clear line that you can toggle around based around benchmarking and capability analysis, I think that's a great starting point that we can then layer on top of.

>> KEVIN LI: I do want to get back to the compute threshold question, but before we do that, I just want to pose a more general question about the policy interventions.

What kinds of policy interventions are best suited toward addressing the marginal harms that we're talking about from open-weight AI?  Are there policy benefits that also could stand to put us in a position where we're particularly benefiting from open weight AI? And I want to give this one to Sayash first.

>> SAYASH KAPOOR: That's a great question.  I think in terms of policy interventions, we must be prepared for a world in which unaligned models exist.  I think that's the bottom line.

We cannot sort of pin all our hopes on the fact that model alignment techniques will be universal and that no one will have access to unaligned models.

So then what do we do?  We need to take a more systemic perspective on how to deal with such risks.  Elizabeth mentioned downstream protections.  I think that's the big one.

And these types of protections, they translate very well to these other risks as well so in terms of cyber defense, it means putting in more investment into cybersecurity, vulnerability detection using LLMs and DNA synthesis tools.

Disinformation, it's incentivizing social media platforms where the disinformation actually spreads to up their game when it comes to detection and removal of this type of disinformation.

I think that's one type of mechanism to detect and prevent all of this harm.

And then when it comes to benefits, I suppose one of the best examples comes from

the U.S. government.

So the U.S. government Office of Management of Budget has been one of the strongest supporters of open source systems in the past, because of a requirement that in many cases, providing the tools, they have to make their models open source.

So any requirements on open sourcing these models would become fair speech. You cannot have a general -- everyone should open source their models, but for the models that we want to scrutinize most heavily, for the models that are used in most critical infrastructure, that's one type of policy lever that can work very well at increasing transparency around these systems and also making sure that we are not using the stick approach too much, we are also offering the carrot.

>> ELIZABETH SEGER: So in terms of what can we do to promote the benefits, I think it's also not just about supporting open source; it's about supporting the whole open source ecosystem which means you know, like we need -- if you want people to participate in open source development, providing compute infrastructure, providing support tools, educational opportunities, I think that's all part of really reaping the benefits from open source ecosystems.

I think when it comes to regulating to prevent risks, I would really emphasize having regulation across the AI value chain so especially in the deliberations around the EU AI Act, one of the main issues was well who holds the regulatory -- who's responsible for the regulatory burden?

Are we regulating the developers? Are we regulating the deployers, and I think the answer needs to be both. Developers of systems need to be responsible for the quality of the systems they're putting out, do safety checks, but you also need the deployers and providers of the systems to make sure that they're putting out quality products.

And when we talk about the AI value chain, we're talking all the way from initial data gathering and data processing and how that's done, making sure data sets aren't biased, all the way through model development and training, model deployment, and then model operation and monitoring, as well.

So having things in place throughout this whole system is what we need to do to keep ourselves safe.

>> KEVIN LI: -- I'm particularly interested in your perspective on realizing the benefits of openness in this space.

>> PETER CIHON: Sure. I like a lot of the comments that I've heard thus far and maybe can transition and talk a little bit in the EU policy making context, which is very informative of having two years on democratic deliberation on topics that we're thinking about.

But before jumping to that, I really do appreciate your points around the government support for open source and Elizabeth, the broader, how can we expand compute resources for underserved actors? I think the national AI research resource is a great proposal along these lines. Generally, as I kind of outlined, one of the key benefits in my mind of open source and openly available model weights is really research, research to understand security, research to understand the safety and improve that for closed or open models.

And kind of expanding eval science I think is going to be a helpful point for open source, and for closed source. The government efforts under the EO are going to be very useful in terms of identifying those benchmarks or capabilities and red teaming and a lot of the evaluation science that we do need.

A key question in my mind is how can we support open source stakeholders with these best practices to implement them in practice?

In order to see who gets the benefit, we need to make sure that we give the tools to open source developers to kind of build securely and implement securely.

But do you want to say something?

>> KEVIN BANKSTON: Yeah, before we dive into the AI Act, although I think this will be relevant to it as well, I do think it's important and relevant that the EO did mention when talking about what to do around open source is that it explicitly mentioned self-regulation as well as regulation. And I think that's important because I think what we expect developers to do and what we think it's appropriate for a government to require or especially the U.S. government to require, likely will be different.

For example, like we have collaboratively, in a scrum of policy, ultimately arrived at norms for companies to follow around trust and safety.

Those are things that we would not let the government dictate because we don't want the State to have disproportionate power over there major platforms that define how we communicate or learn things.

That's just good policy. It's also a First Amendment value.

That doesn't mean that we don't need to use every tool in our toolbox to push developers to invest as much as responsible in safety in proportion to how much they're investing in capability.

So, but I think that means we need to be very clear to them and to ourselves about which problems we are targeting, and make sure that whatever interventions governments are proposing are narrowly tailored to address and are actually likely to address those problems.

One because that's just good policy.  Two because it's probably required by the First Amendment.

>> KEVIN LI: Did you have a follow-up to that?

>> PETER CIHON: I did want to pick up on the thread that Elizabeth mentioned around advocating liability and responsibility in the value chain, drawing on the AI Act if that's already.

GitHub has been quite involved in EU policy making as kind of European folks are looking at taking the product liability and product regulation scheme in the CE marks that we're all aware of to the digital world and that's in the AI Act, that's in the Cyber Resiliency Act which closed in November.

And all three of these files have addressed open source and the question of upstream downstream, and I think it's helpful to see what those outcomes are and what we might learn in the U.S. context.

So as you rightfully called out, Elizabeth, there are particular cases where open source is treated the same as closed source and the largest models, ten to the 25 flops of that training are given the presumption regardless of whether they're open or closed that those need to be within the scope of the most stringent expectations on upstream models.

That's an exception to the larger kind of norm that we've seen in these three files, which is really a clear distinction between upstream development and downstream deployment and there's a huge focus or the liability and the expectation to comply with the rules across these three files is on the downstream.  It's on the folks who are integrating components that they might find on the internet for free from open source contributors, whether that be a company or an individual and they're integrating that into a digital product, into an AI system.

They're then offering that to customers and the expectation is that they're going to make sure that customers are not harmed.  So the liability is really falling on folks downstream and that's having knock-on benefits where the folks upstream are continuing to be free to innovate, you can have a graduate student who's able to put out a model that might be incredibly useful for folks downstream, but not face the burdens of compliance because they're offering this for free.

And we're seeing folks downstream incentivized to choose the correct package.  So if you are thinking about integrating some type of software component, you can pick the one that's maintained actively and has certification and clearly visible on GitHub or you hopefully choose that as opposed to something that was put up over the wall as a final project for a Ph.D. student.

And so that norm has been kind of established or recognized and maintained in these three files in the EU context, and there are exceptions, and I think that it's the risk-based approach to finding where we need to draw those exceptions.

Open source systems in the AI Act, we haven't seen the final text, we have the political agreement, we've seen some leaks so we need to caveat it, of course. High-risk systems, you're in scope. You shouldn't be able to get around this idea of you find a loophole because simply the system is open source.

And same with systems that face obligations for transparency. So someone will not lawfully be able to take a model, implement it into a system that purports to be a human and lies to the end user, regardless of open source or closed source.

And so I think it's important to acknowledge there are exceptions to kind of this carte blanche for open source, but this distinction between upstream developers who are enabling folks to do many things and the downstream folks who are integrating them into commercial products that affect people's lives directly are facing the burdens in supporting openness and there are lessons for us to draw in the U.S.

>> KEVIN LI: We could talk about the EU AI Act for weeks if we got the chance. I want to return very briefly, at least to the question of compute thresholds that the EU AI Act raises and also raises the AI EO here, and I was a little surprised by how positive the sentiment around weights were.

But I want to open the floor for a moment for folks to give -- not about weights, about compute. I want to open the floor for a moment for folks to talk about the compute thresholds and what might be positive and negative uses of a compute threshold when we're thinking about risks, especially in open weight models.

>> SAYASH KAPOOR: I can start.

So I think if we look at what the actual risks of open foundation models have been so far, presumably the biggest risk as Kevin alluded to earlier has been the risk of nonconsensual pornography. We have NCII that has been enabled spreading across the internet, in particular in -- it's CSAM, which is spread across the internet and, in fact, the national center for missing and exploited children reported a sharp uptick in the amount of content they've received.

A lot of it was attributed to open source AI systems. So I think one of the biggest challenges with open foundation models is exemplified by this spread because in this case it really was the hobbyists who were removing the safeguards and generating nudity or pornography and putting this content on the internet.

What do we do about it?

It turns out the models can run on your iPhone. They can be trained using orders of

magnitude lesser money compared to the state-of-the-art frontier models. Anyone can download the system on your funny and use it to generate NCII.

And I think that presents a real issue and it's really problematic, it problematizes the issue of compute thresholds in the first place because the biggest risks require AI models that are not really the frontier. They're not even close to the frontier. They are things that can be run on your Mac book or iPhone.

So this points to the limitations of this compute threshold as a way to manage risk and it also points to the limitations of the nonproliferation regime in the first place because we cannot expect to make sure that these models don't proliferate because malicious users can train their own.

The models that have been already been released so far, are already good enough to create NCII and CSAM. This seems like an issue because on the one hand we are talking about what risks of AI we want to mitigate. The biggest risk to mitigate would not be mitigated by compute thresholds and it would not be mitigated by nonproliferation. This points to the downstream interventions where we can actually sort of look at the attack surface where this harm actually happens.

And we've seen examples of this working out. One of the bigger platforms where NCII content was being shared was a community forum for people sharing AI-generated images where a lot of NCII was shared and over the last two weeks or so, the platform in response to backlash from users as well as the community has for the first time started taking efforts to address the problem of NCII.

This shows, even though we cannot fully resolve the problem, one of the best things we can do today is focus on the downstream harm, on the attack surfaces where such harm actually happens, and to make sure that the existing social contract that we have where companies are responsible for downstream harm on their platforms, even if protected by Section 230, we should continue to push for such interventions as demonstrated.

>> ELIZABETH SEGER: I want to add. I think it comes back to this idea of different models, different risks. I think you're completely right. There's a whole bunch of risks that will not be covered just by having this ten to the 25 flops cut-off and like you said, this is the sexual abuse material, lots of stuff around image generation, a lot of biorisk implications.

And I think in these spaces you're right. We need to focus on the downstream harms. I wouldn't say that that means that having this compute threshold is a bad idea. I think the compute threshold is important for a different class of risks, the far future risks, the things people are worried about when they talk about loss of control risks, highly persuasive AI. These are the things we're seeing coming down the line and having this

compute threshold and watching that frontier is equally as important.

There's been a lot of tension and a lot of discussion between do we need to care about AI ethics issues or the more knock-on AI safety issues, very often wrapped up in issues around existential risk and extreme risk.

The answer needs to be both. We cannot focus on one or the other. There are some extremely important risks, probably the most prevalent risks being realized right now that are around bias and discrimination and harmful material being shared online.

Just because we haven't seen more extreme risks come into play doesn't mean those aren't things we should be looking out for.

And when you're looking at that threshold, when you're looking at what the more highly capable models down the line could do, having that compute threshold is a really nice cut-off for being able to say let's watch what the frontier is doing. Does that mean that the most dangerous systems are necessarily going to be above that threshold? Maybe not. We're seeing amazing things being done with compute efficiency and algorithmic efficiency.

We might get some pretty amazing capabilities below that compute threshold so I think that means we need to be flexible. We need to reevaluate the thresholds and also, have the flexibility to sort of move models between tiers, depending on the capabilities that we're evaluating.

>> KEVIN BANKSTON: I just want to second that the sort of polarization between the catastrophic risk and the more -- I hate to say mundane risks, but immediate risks.

>> PETER CIHON: Immediate harms happening today.

>> KEVIN BANKSTON: -- is not helpful, especially because so much of the mitigation work that needs to be done will be the same, regardless.

Open AI is very catastrophic focused and doing an enormous amount to mitigate immediate harms and risks, a greater proportion in terms of their investment in money and people than the biggest companies. We would like to see those percentages change at the bigger companies, too.

Where I differ I think or where I think we're all going to need have a conversation to get to the bottom of it is if we're all agreed that the likely risks that we see including chemical, biological and nuclear are likely not going to be prevented by a cap, then what risks specifically are we trying to prevent? And if they are very speculative, speaking generally, we have not regulated super speculative risk and we haven't done that when it could impact access to what we're all describing as the most useful general purpose tool for knowledge and expression we've ever seen.

And so I think really -- this is not saying we should not worry about these. I think this is agreeing that we need much more research to much more concretely define what are the risks we expect from the largest models? Right now, I think it's very unclear.

>> ELIZABETH SEGER: I think much more research is the key starting point here.

But I also think we need to realize that we're not regulating just because we aren't seeing evidence of the harms, is maybe not the best option, either.

Let's be prepared. Let's have procedures, policies in place for the "if" scenario.

So this isn't saying necessarily like regulate the systems we have now, but be prepared to regulate what might come down the line, and identifying what some of those thresholds are.

So I think that's where we need to focus.

>> PETER CIHON: Another reason to love the EO from my perspective is that they are kind of charting this middle ground. It's about monitoring models at this middle level. It's not about mandating, regulating, but reporting to the government about what kind of practices are being done and risks are being found.

And I would agree that there is need for caution at the frontier. There's been in the community a motivated -- ongoing assumption I would say that there are emergent capabilities that we fail to predict and that is motivating a lot of policy making. I think it does warrant caution, but it's also worth noting that this week, in New Orleans, the paper that was out of Stanford about six months ago, are emergent abilities a mirage won the outstanding paper award. A clear signal from the research community that we need better evaluation science to be able to predict capabilities so that we can move to a more informed future --

>> KEVIN BANKSTON: Can you let the audience understand why that was important?

>> PETER CIHON: You were giving a great download on this in the hallways. I would look to you if you don't mind. I'm happy to give an attempt.

>> SAYASH KAPOOR: So I think the key contribution of this paper was just to say that a lot of claims about emergent capabilities in models actually come down to the metric that you choose to evaluate them.

So seen one way, the emergent capabilities, but another way, they're gradual progressions and in that sense, they're extremely predictable.

>> KEVIN LI: Before we move to audience questions, I want to return one last time to the question of benefits and especially in research and innovation, which is something that we've been talking about a lot here today.

And I think it's easy for research and innovation to be -- and, you know, related concepts to innovation like competition to be relatively abstract questions of the more open the model, the more research innovation you get. Can you offer concrete examples or thoughts on how, especially open models might contribute or supercharge the ability for researchers and innovators to use AI in ways that we might want them to? And that's for anyone.

>> SAYASH KAPOOR: I can take a stab at it.

So I think like it cannot be overstated how important open models have been in the last two years or so, in terms of research.

So as Peter and Elizabeth have mentioned, we don't really understand a lot of the reasons why -- behind why language models work the way they do, and I think that's sort of the clearest trend. Here are a couple of concrete examples to Kevin's point.

The first is this work out of CMU which basically showed that you can add adversarial strings to bypass safety filters even for closed models.

This is research that would not be possible. It's impossible to do this research without access to a language model because it requires looking at the weights and the gradients and so on.

And so this is research that we can directly pinpoint towards and say this research was enabled by open access to models.

There is a plethora of such research. So another strand of research is on verifying what data was used to train a language model. And in particular this type of research might be useful if it comes down to liability issues for copyright protection and so on.

And if it comes to verifying whether a language model was trained on a particular data set or not, we actually did not have any way of doing that until this paper came out six months back.

And so this is one other sort of type of innovation that is being enabled. I would say that all of these research advances require access to differing levels of assets.

The last paper that I mentioned requires access to the data, the code as well as the model gradients, and so this also highlights the gradient of open release that we talked about earlier.

The further along you are on the gradient, the more visibility you have into sort of the data sets used the train the models, the code as well as the models themselves, the easier it becomes to do this type of research.

In general, a lot of the research has been enabled by the extreme ends of the gradient and in particular, for instance, nonprofits which are released their models as well as the

code and the checkpoints of the models like what happened to the models while they were being trained openly.

And I think this really sort of underscores the importance of open source AI, where you have access to the entire sort of chain of materials that was required to prepare a model.

>> ELIZABETH SEGER: I might comment on market concentration, as well.

So one strong argument in favor of open source is we have open source models. It allows more people to learn from models, to use them to employ them downstream to build on top of them.

You have many more people joining and getting involved in the AI market space and that's huge for disseminating profits away from a few key players, allowing more people to benefit not just from the use of AI, but being able to develop much more diverse range of technologies.

I think there is a lot of uncertainty, though, and an open question around how important open sourcing highly capable frontier models is going to be for market concentration or making healthy market concentration.

And this has to do just with the cost and the amount of compute that it takes to train these extremely large models.

For example, I think it's projected that training GPT5 might cost somewhere in the region of $10 billion or something like that. It's an absurd amount of funding.

So I think when we talk about open sourcing these large models, we're not 100% that those are necessarily going to be the most commercially viable systems. We might have a lot of the capabilities that these large models have in much smaller models and open sourcing the smaller models will do the vast majority of the work for getting lots of people involved and be able to benefit from this open space.

And if that's the case then open sourcing these larger models might not have the most significant impact on market concentration.

We don't know that, though. There's a lot of uncertainty. We don't know how this is going to unfold. A lot more research is needed to understand where these capabilities are going to go.

If it is the case that some of these more frontier models have extreme leaps in capabilities and those are the most commercially viable systems, then the impact of open sourcing these models on market concentration will be larger.

So I think that's something to focus on, as well. There's a lot of arguments around market concentration, but there's a lot of uncertainty around how models are developed,

where the capabilities are going to be and how those capabilities balance with the costs of developing the models in the first place, which leaves just a huge question mark over the -- what are the market implications of open source?

>> KEVIN BANKSTON: I agree with a lot of that. I think if we talk about the large expense, one that does support the goal of having more of that be a public good than a company secret.

But also, we've seen the release of the larger models has led to research, has enabled research that led to those smaller models where they're making exponential leaps in how can we make this much smaller while providing much of the same value?

And not only is that I think important knock-on benefits, but it also points the way toward a less centralized future internet where, for example, you could have most of your AI on your phone and only call to the big, big models probably never if you're just a civilian trying to do normal stuff.

And in that way, privacy is better, competition is better. You don't have single points of security failure in these centralized companies.

Everyone says we can't make the same mistakes we made with social networks with AI.

I would say if we disincentivized centralization of the power, we are replicating the problem with social media and we ought to think more about how do we ensure dispersion and decentralization of that power?

Because I know you see the social network market trying to build those models now as a replacement for the centralized models. I would rather build them first.

>> ELIZABETH SEGER: I think this is a great point to come back to this idea that there's a false dichotomy between completely open and completely closed and that there are genuine risks associated with widely sharing weights, especially for more highly capable models but there are also very genuine benefits in terms of model development, competition, and if we're going to make progress and find this healthy balance, we need to do away with this open versus closed dichotomy.

There are so many options between fully opened and fully closed that we can pursue for research purposes, for downstream development purposes.

They don't have necessarily all the benefits of open source, but they also don't have all the risks, and I think part of this is going to be exploring some of those middle release options and seeing how can we get the most amount of benefit from them while preventing risks?

And then I think on top of that also understanding that there are often alternative options that we can use for pursuing some of the benefits traditionally associated with open

source.

So when we think the risks of open sourcing might be extremely high, we might be able to engage in other activities to try to get some of those benefits, as well.

So, for example, for transparency, we could encourage impact statements to be published, information sharing with regulators. For distributing profits we can tend to windfall clauses or taxation redistribution schemes.

These aren't straightforward options, but this idea of where the risks are high, if there's a will for pursuing these benefits, there is a way.

So I think the main thing I want to get across here, though, is let's get away from these dichotomies of open versus closed, always open versus always closed, and the idea that if you're one group or the other, that you're not saying all models should be closed or all models should be open.

There's a lot of room in between to explore.

>> KEVIN BANKSTON: Agreed and I'll point folks to a policy paper that Sayash and his colleagues put out just today that has a really great discussion of the different spectrum of responsible release approaches, building on work and maybe someone will ask questions about the spectrum.

>> KEVIN LI: I expect gradient is a word we'll be hearing quite a lot in the coming months.

I want to open it up for questions and I know there are questions coming in both from the internet and from folks in the room, but if you're in the room with questions, feel free. I think everyone has index cards on your seat hopefully, feel free to pass those over to your left and in the meantime, I think these might be questions from the internet.

So the first question is how do we decide what risk level we can accept as a society as a trade-off for the benefits that we've been talking about today? It's a hard one.

>> ELIZABETH SEGER: It's a very difficult question. I would say that we need to have democratic processes central to this decision making.

So this is a question -- we're going to have lots of cultural variation, people from all different geographic locations, walks of life.

So I think implementing democratic processes to make decisions around accessible thresholds is going to be key. There are some groups doing amazing work on this.

So the Collective Intelligence Project, they're doing great work in this area and I would make my bid for democracy here.

>> KEVIN BANKSTON: I'll just add. We don't know the answer to that question, and I

think we've left it unanswered in a lot of key places because right now, we talk a lot about introducing risk management systems, the EU AI Act is going to require it for high-risk systems.

And we often stick at the issue of what is the taxonomy of risks and what are the potential mitigations? We rarely talk about okay once you've identified those things, how is a company supposed to make that decision? How do they weigh those equities? Or is it just paperwork for them to articulate the risk and the most senior executive signs off on that risk, which is not very helpful.

This is something that our Governance Lab is going to be trying to dig into is how do we systematize those systems that isn't just more compliance paperwork that doesn't make people safer.

>> KEVIN LI: It might be -- since I have you all here anyway, it might be easier to flip back and forth between the index cards and people in the room. So I'll take a question. Yes, please.

>> I was interested in the panel's view on what level of testing companies should put their model through before releasing them openly, given how much uncertainty and disagreement there is on what to accept from next-generation models and what the risk would be and once they're released, there's no putting the genie back in the bottle. For different kinds of models, the frontier models -- (inaudible).

>> KEVIN LI: I'll repeat the question for the folks watching because I know the mics don't reach back there.

What kinds of testing should companies engage in before releasing these open weight models? Because and in in particular once the model is released, there's no unreleasing the model.

>> ELIZABETH SEGER: I'm a big fan of staged release as an option. So this is a process by which you basically release a smaller version of the model behind an API so it's still reversible if bad things happen. You study how it's working, what are the societal impacts? How are people using it? How are people trying to misuse it? And then you implement safety filters and so on.

And then you release the next bigger iteration and the next bigger iteration, more capable, more capable and taking a break at each step to understand how is it being used? How is it being misused? What are the societal impacts? And if everything is looking good, then I guess go ahead and open it up.

One nice thing about this process is in the process of staged release, if you're noticing that you're layering on safety filters and implementing lots of fixes, that would be a good indication that if you were to open source that model, that those would all be things that

could easily be removed to allow the model to be misused, and then that would be a risk assessment that you have to do and decide whether it's still worth it at that point.

But I'm a big fan of staged release.

>> PETER CIHON: Evaluation science is not the science that we need today. Information and knowledge about how to do this well is not evenly distributed. It's in a few leading labs, and it's really encouraging that the Executive Order is mandating NIST to go and identify best practices that can take some of that information and broadly disseminate it to society so more developers have those state-of-the-art evaluation methodologies without kind of espousing a particular view of whether they need to stage release that or not.

But the idea of performing appropriate evaluations before getting anywhere near the decision to release is one.

The EU AI Act, the text is shifting as we called out already, but at that level of the general purpose models with systemic risk, there's an acknowledgment expressly of adversarial testing, as well.

Again, not having any kind of bearing of whether you're not allowed to release the model or not, that's not in the legislation.

Instead, it's about these are the types of processes you should do.

>> KEVIN BANKSTON: I think red teaming does and should imply external and internal red teaming. That's bringing voices in and making sure that's a diverse set of stakeholders so you don't have the four of us telling you what to worry about but have affected users and people who have different life experiences actually highlighting what they're most worried about.

>> KEVIN LI: Our next question is from the internet. The question is, place yourselves in the shoes of a founder and perhaps working downstream of these open weight models. If you wanted to be responsible, what are the considerations given limited resources or limited budget? What are the top priorities for making sure that your deployment of these models is safe and rights respecting?

>> KEVIN BANKSTON: It highly depends on the context of what service you're offering. It's going to diverge. And this is one of the benefits of fine-tuning which you can do on Open AI's server or on top of an open model because the kind of harms you want to guard against in a children's social app are going to be different from the ones you want to guard against in a system that decides how long someone gets parole or what public benefits they get or what medical care they get.

>> ELIZABETH SEGER: I would add on as well, one of the things being discussed in the EU AI Act was where the compliance costs are going to land. And so I think one

thing to keep in mind, if you're a downstream developer you might not necessarily have the resources, if you're building on top of a very large system, that one of those megacompanies put out, you might not have the resources or the know-how in-house to do the kinds of safety checks and stuff that might be ledge slated.

This is an argument for coming back to this regulation through the value chain thing making sure that developers of these systems are also doing -- broader but broader safety checks and evaluations before downstream developers get their hands on them, so that it's not all on the downstream developers, all the compliance costs, all the regulatory burden for making sure that deployments of these systems that work in ways unknown to us don't fall on the smaller developers.

So making sure that those compliance costs are split between the upstream developers and the downstream developers.

>> PETER CIHON: Setting aside compliance costs and the future regulatory approaches, it's useful just to take a look on GitHub or Hugging Face and see is there a model card that gives you assurance and -- and gives you the information you need to put it into practice, taking expertise from folks in the deployed case in particular because it does vary to your point and do so safely.

So look at documentation first and foremost. There's a lot of information that's out there, a lot of best practices.

>> KEVIN BANKSTON: And I think norms around that documentation are going to be really important, not least because the EU AI Act requires it, but gives very little detail on what it should look like. I think that's a really important point of intervention including for NTIA and NIST as they think about standards, just as NIST's work on risk management works has been and will be really important in sort of populating the thinking of the EU AI Act's mandate for such systems.

>> KEVIN LI: I think we've got time for maybe two more questions, and so yes, you in the back, sir.

>> So Elizabeth, you mentioned the assessed risks will be taken into account in different contexts -- (inaudible) not all countries express democracy the same way we do here. So what is the role of thinking about open source when not all players in this space will be using democracy to make decisions about how --

>> KEVIN LI: I'll repeat the question. How should we be thinking about these decisions when not all of the relevant players share the same values, for example, of democracy? Elizabeth?

>> ELIZABETH SEGER: That is a very tough question, and in some ways the question. I don't know if I'll give a great satisfactory answer here.

I think one thing we have going for us is right now, most cutting-edge frontier model development is happening in a handful of countries. So when it comes to the regulation of those models and making decisions about risk assessment around those, the regulation in those countries, mainly democratic countries right now, is going to be quite central.

Open sourcing is important for distributing the use and benefits of these systems very widely. Yeah, I think one thing that the EU AI Act does well is if we have legislation like this that has stipulations for best practice for major AI developers within the jurisdiction of that regulation, that will in a way kind of -- trickle down is a bad terminology, but it will have an impact in other countries where those systems are deployed and also give those other countries recourse for being able to claim against harms from the systems based on the regulation in other countries.

So I think if anything, that shows that the regulation in the EU AI Act and the regulation that we're considering here in the United States right now, this is standard setting, not just for our countries, but global impacts, as well.

It's a tough question.

>> KEVIN BANKSTON: You do raise a tough question. It ties into the marginal risk question. Is this a regulation that will ultimately not be helpful if any significant actor, bit torrents, highly capable -- on the global internet, and is happening from a variety of countries, some of which we don't agree with on a lot of things.

>> SAYASH KAPOOR: And I think this goes back to the point that we cannot rely on alignment of individual models as a mechanism for systemic safety.

Unaligned models are going to exist. They already do exist and we cannot rely on the fact that a single model or set of models even would comply with these regulations as a means of curbing their misuse. We have to rely on things that are more robust, downstream protections and so on and this comes back to the point of compute thresholds in some sense. While ten to the 26 or 25 or whatever number of flops you want to regulate might seem like a huge number today, the compute costs for creating these models goes down exponentially over time.

Over a span of seven years, the costs of creating an image model with the same amount of capabilities went down 44 times. And this process is only improving and increasing as we invest more into AI-specific chips and AI-specific architecture improvements and so on.

So all that is to say any sort of concrete number we come up with is likely to be meaningless five years from now because a group of gamers in a LAN party will be able to hook up their gaming laptops next to each other and train the same model.

>> KEVIN LI: And perhaps one final question from the room? Russ.

>> A very self-interested question. Also on the democracy point. So effective democratic self-governance requires some degree of expertise, knowledge, basic understanding. I think it's also the case that AI policy issues will be best addressed in part by a population that is savvy about misinformation, about deep fakes and bias and so on.

What can policy makers do to help facilitate and engender an AI-savvy, AI-literate population?

>> KEVIN LI: The question is given the importance of the expertise that you all bring to the table, what can tech policy makers do to foster that conversation among the population as a whole?

>> KEVIN BANKSTON: Talk to CDT.

[Laughter]

>> KEVIN BANKSTON: That wasn't a joke.

>> ELIZABETH SEGER: I think that there's a lot to say for educational campaigns more widely. Having people interact with the systems is also huge. This is what we saw with the release of Chat GPT. AI went from a thing that AI researchers talk about to something the whole world was talking about. Suddenly my grandmother was asking me about my job and that was weird.

Having interaction with the technology, nice user interfaces, people can experience the systems. If you're able to create images yourselves, that opens your mind to the great things the system can do, but also potentially the nefarious uses, as well.

So I think first-hand experience is huge.

And then in the process of actually making policy, making these decisions, have lots of stakeholder engagement opportunities for bringing people into the room. Have these discussions.

And that would be the way forward on that. I think from a base population in general how do we create, let them experience.

>> SAYASH KAPOOR: To that point, I completely agree and I'm not sure how many of you saw the image of the Pope, I think that was a great example of not really an example of misinformation, it was a meme that went viral and through that meme a lot of people realized what the capabilities of these models are. That's a prime example of insulation against these types of concerns, proliferating to the general audience.

>> PETER CIHON: Societal inoculation via meme. I like that quite a bit.

I want to plus one the point on education. We can have great public education components. CISA is running one on social media best practices and hygiene, we might imagine one in the future around AI disinformation.

There's a plug for open source in general. This is a great opportunity to get small models in the hands of computer science students. We can push and increase computer science education broadly I think is going to be very important and enabling more folks to build and tinker with these models. They might go on to careers that are not in certain large companies and get to work for the government. So I think that's an important piece in the medium term and to Elizabeth's great point around open consultation, seeing a lot of stakeholders participate, I think that it's really encouraging that the Biden Administration went the route of consulting to get widespread input on this challenging question of widely available model weights.

I would implore you to listen to folks that are building and using these tools day in, day out to see what they can do when they're given that ability to tinker with model weights directly.

>> KEVIN BANKSTON: One addition, which is -- well, two. One on a lighter note. I would advise policy makers to watch less science fiction in the sense of -- actually -- watch science fiction, I love science fiction.

But I do think it is fair to say that the portrayals of AI in science fiction over the past several decades have somewhat polluted the discourse in how we talk about it now and how civilians who aren't familiar with the technology think about it.

And like pictures of the Terminator don't help.

On a more serious note, I think more technical expertise in Congress is critical. I would highlight especially the great work of Tech Congress, which puts tech-literate fellows on the hill, but we need more systemic approaches to better educate our policy makers and staffing them with technical experts.

>> KEVIN LI: Well, thank you very much and on that note, I have one final question and it's a good one. It's from the live stream. We've done that one. What opportunities are there to contribute to the Executive Order or NTIA policy making on this topic? Which I can take.

[Laughter]

As assistant secretary Davidson said, this is the first step in our consultation process. We hope to do more public outreach. We'll for sure be doing a public request for comment as directed by the Executive Order and we encourage everyone to make your voice heard in that proceeding.

We are really privileged and excited to be able to engage with so many brilliant people

across the policy-making spectrum and across different stakeholder groups and, you know, we really look forward to hearing from you in person, over the internet.

We have an open door policy so please reach out.

With that, I really want to thank CDT and our wonderful panel here for being both such great hosts and such great panelists.

And hope to see you soon.

[Applause]

[Event concluded at 3:29 p.m. Eastern Time]

* * * * *