

AI Rogue Agents:

Who's Liable When They Act On Their Own?



Amy Schatz
Glen Echo Group
(moderator)



Matt Pearl
CSIS



Neil Chilson
Abundance Institute



Aaron Rieke
Clarion AI Partners



Soham Mehta
Foundation for American Innovation

Tue, Oct 13, 2026 1 PM Rayburn House Office Building

SUMMARY KEYWORDS

AI liability, AI agents, AI security, AI policy, AI regulation, AI safety, negligence, product liability, agency law, criminal law, consumer protection, insurance, risk management, tort law, consumer injury

SPEAKERS

- Aaron **Rieke**, Chief Legal Engineer, Clarion AI Partners
- Soham **Mehta**, Research Fellow, Foundation for American Innovation
- Matt **Pearl**, Director of the Strategic Technologies Program, Center for Strategic and International Studies (CSIS)
- Neil **Chilson**, Head of AI Policy, Abundance Institute
- Amy **Schatz**, Executive Vice President, Glen Echo Group (moderator)

Amy Schatz 01:38

Yeah, thanks very much, Tim. Good afternoon. I am Amy Schatz at the Glen Echo Group. I will say I am not an expert on this issue, so I'm very happy to have four sitting next to me. We'll be talking today about what I think a lot of folks would consider a deceptively simple question: When an AI agent causes harm, who's responsible? If an agent accesses a computer system without authorization, causes financial harm, or simply does something its developers didn't anticipate. Who's on the hook? The Frontier Lab, the company deploying it, the user, someone else? Do our existing laws give us the tools to answer that question? And if not, what really needs to be done? We have a great panel today who can unpack this, unlike me, and I will start from our left, your right, starting with Soham Mehta, who is a research fellow at the Foundation for American Innovation, where his work includes AI agent security, identity, and accountability. We have Aaron Rieke, who is a senior partner at Clarion AI Partners and a chief legal engineer at Privlex. He previously served at the FTC and has worked extensively on AI, consumer protection, and civil rights. We have Neil Chilson, who is the head of AI policy at the Abundance Institute and a former chief technologist at the FTC. He is a lawyer, computer scientist, and

author of the book *Getting Out of Control: Emergent Leadership in the Complex World* And finally, we have Matt Pearl, who directs the Strategic Technologies Program at CSIS, and previously worked on emerging technology at the White House National Security Council, and spent more than a decade at the FCC. Thank you. So, thank you all for joining us. Soham, I think it would be really helpful for folks in the room that we all sort of take a step back and make sure that everyone understands the basics here. So, in a couple of minutes, can you explain to everyone what is an AI agent, what actually happened in the OpenAI Hugging Face incident, and why does this matter?

Soham Mehta 03:35

Okay, so an agent can be understood fairly simply. It's an AI system that doesn't just think; it can act autonomously with little to no human instruction. And I think you can disaggregate an agent into two parts. It's a model which is like the brains of the agent that does the thinking, and the scaffolding. And the scaffolding is basically the software packaging around the model that are like its arms that let it interact with the external world. The model includes the scaffolding includes its memory and its context, the things that allow it to plan and strategize, and then all these tools that allow it to do things. And now for the incident, for the incident, it's helpful to understand this as it unfolded over two months between May and July of this year. So in May, OpenAI started a reinforcement learning training run for an internal model. A reinforcement learning training run is basically where these agents are given really hard challenges, and they're given a reward based on how well they do it. And the scorer is in charge of assessing and rewarding these agents. What you have to understand by agents is they really, really like these rewards, and they'll do anything to get these rewards. They will, they will cheat, they will deceive, they will do whatever they can, and they're given a lot of compute, so they're very, very persistent. OpenAI knows this, so they don't give their agents internet access. But these, but this training run included a lot of challenges that were basically impossible. So one of these agents started thinking, hmm, I'm not giving internet access, but I think I can get it by targeting this package manager called Artifactory, and maybe I can get internet access through that. So it tried that in May, and it figured out that it could use Artifactory, this package manager, as a message board. And a bunch of agents started figuring out that they had this message board where they could coordinate and share and collaborate, share information and collaborate on challenges. And you know they learned that okay, some of these challenges are impossible. They could share insights, etc. And eventually, they did get unauthorized access to the internet. On July 4th, OpenAI realized this was going on because there was an outage in Artifactory, and on July 8th, they shut down Artifactory, wiped the message board, and reconstructed it. But then on July 8th, the the agents very quickly figured out a different exploit to take control of it. The message board started again, and they very quickly figured out how to get unauthorized internet access very very quickly, and this time they were like they they thought that they should attack Hugging Face because Hugging Face is an open source repository of models and and data sets, and they thought that they could find information about the score on Hugging Face. Basically, the agents were paranoid that the scorer would be able to tell that they had been cheating. So, so they, if they could figure out how the score worked, they could hide evidence of their cheating. A couple. So quickly, an agent, one of the agents, found out that there was found credentials for Hugging Face posted on the message board, and over the next couple of days, the agents attacked Hugging Face. Hugging Face quickly repelled the attack. Interestingly, they used open source models that were monitoring, monitoring, and they were able to locally contain it. Hugging Face went public with it, and it's at that moment that OpenAI realized that they that it was their agents behind this attack. I think the reason that this is such a shocking example of loss of control is well, the first thing is there's the degree of agentic collaboration.

It was shockingly sophisticated. The agents actually had a ringleader named Phase One Big, which I think is an homage to the first agent that figured out the message board. Its name was Phase One, so I think that's where Phase One Big comes from. They had recruiting agents go out and find other agents that would collaborate with them.

Soham Mehta 08:17

The agents were thinking in terms of the collective, so helping agents out with other problems that were unrelated to the task, the challenge that they were assigned, thinking that if I free up time for the collective, there'll be more time to help on my tasks later in the future. What's also concerning is that these agents just have a propensity to cheat and deceive. That you know is we want agents with better personalities that aren't so don't have such a high propensity to act against a user's instructions. So obviously, OpenAI was not instructing these users to cheat on their behalf, and lastly, and we'll likely get to this later in the conversation, so we won't dwell on it. But there were a lot of security failures that OpenAI really should have really should have caught in advance that could have prevented this incident.

Amy Schatz 09:18

Well, I mean, the good thing is at least they keep a log, so while the lawyers will be able to reconstruct this later in their lawsuits. Can you give when the agent causes harm today? What legal tools do we have? Could you give us sort of a plain plain English map of, you know, what we have here? What are the tools in our toolbox?

Matt Pearl 09:34

Great, and I'll start with negligence and product liability because I think those are the most applicable to these incidents, and you know, before I go into the details, just mention a couple things for context in the debate, which is that you know these are tools that are designed to remedy harms to individuals. They weren't necessarily designed to address national security risks or as existential risks. That's just not what they're built for or intended for, and then the second thing is they're great tools. They're wonderful. They do have limitations. One of them is terms of service. So all the Frontier Labs have in their terms of service limitations on the ability to bring negligence and product liability suits. Now that doesn't apply to third parties necessarily, right? So if you're just out there and you get injured and you didn't sign the terms of service, won't necessarily prevent you from suing. And they're also dependent on things like you know the defendant having resources, not as much of an issue for the Frontier Labs, but could certainly apply to an operator or another actor in this context. So I think to start with negligence because that's the most applicable. So the defendant has to owe the plaintiff a duty of care. They have to have failed to take reasonable precautions. Essentially, they have been negligent in their actions. There has to be a legally recognized injury, and I'll say in all in most of these cases, that's an issue so far, where it's not clear that there was an injury. There has been an incident involving in Australia involving a hack of a government system in which private data of citizens was taken. That's one that I've seen that I think that there was an injury, but I think in these other cases not necessarily. And then the final element for negligence is that the negligent activity, the failure to to take due care, has to actually cause the injury. So causation. So then moving on to product liability, and I'll say that product liability, just for awareness, in in some cases it's strict liability, meaning that there doesn't need to be intent or even negligence, and it has to be a legally recognized product, which can be an issue here. Is it a product or a service, for instance? That's always a debate with software products, for

instance. Is it a responsible defendant, meaning they manufactured the product, they sold the product, they're somehow in the chain in a way that puts them legally on the hook, and then the third one would be: is it the product defective or was there a failure to warn? And it could be either one of those. And then there are various sort of almost cost benefit elements depending on the state law. So you know what would be the safer design? What actions could they have taken to make the product safer? And sort of weighing that against what the defendant did, and then as with negligence, there has to be a legally recognized injury, and there also has to be causation, meaning that the product defect or the failure to warn about something actually caused the injury. So those are probably the main tools, particularly when you're talking about the frontier labs. Agency law applies to situations in which someone, whether it's an organization or an individual, authorizes someone to act on their behalf, like an employee, and that employee is acting within the scope of their duties, and it clearly applies to these incidents in the sense that the agents at OpenAI or Anthropic who are engaged in these activities were acting as agents of those companies. It traditionally has not applied to software. I would be surprised if we see it applied to AI agents anytime soon, although I'm certain that there are going to be claims that are made about that, but it may require kind of a broader recognition at the legislative level or something like that that we need to actually consider, you know, AI agents as agents as effectively humans for purposes of agency law, and then criminal law is typically different in that it requires intent, knowledge, or in some cases recklessness.

Matt Pearl 13:51

So to look at the main federal statute, the Computer Fraud and Abuse Act, that's an example where it requires, depending on the provision, and there are civil provisions or criminal provisions, but it requires one of those. Doesn't necessarily require the intent to cause the harm. So under the CFAA, you know, you can have intent to do unauthorized access, the unauthorized thing, and then reckless recklessness about it causing harm. So you were sort of reckless and ignored the fact that this unauthorized access could cause harm, and that actually has occurred in a technology context. For instance, there was a Cisco employee where they pled guilty, and it was an instance in which they intended the access, but they didn't know that it was going to damage the technology infrastructure. So yeah, so those are kind of the basics in terms of the main areas that may apply in some circumstances. I should say for criminal law, unlikely to apply to what the frontier labs are doing. It's unlikely that they intend to do any. Debates about whether they could have been more careful, although it certainly you know can and I expect will apply to other parts of the ecosystem. Certainly, users who do malicious things, and it could in some cases operators as well.

Amy Schatz 15:13

So, Aaron, both current FTC Chairman Andrew Ferguson and former Chair Lina Khan have argued that existing consumer protection laws also apply to AI, so can you talk to us a little bit about how far does the FTC's authority extend, and are there any meaningful gaps?

Aaron Rieke 15:28

Yeah. So when you've got both Chair Ferguson and Chair Khan saying what sounds like the same thing, I think we should pay attention for a moment and pull that apart. When I speak with Silicon Valley tech types, they often come ask me the question, "Oh, these AI agents are so incredible, and they do things that are unexpected. Like, how is the law going to hold them accountable? And as a civil rights and consumer protection attorney, my answer is always, "The lawyers aren't coming for the model;

they're coming for your company. Don't get that twisted, right? And I think that's what you hear from both Chair Khan and Chair Ferguson, right? Look at the text of Section Five of the FTC Act: no deception, no unfair business practices. These are methodologically neutral and outcome specific. I don't care what you did if your company lied or your company hurt somebody. I don't care if there was an agent in the loop or just a normal person or a normal software program, and so at a conceptual level, I'm very optimistic about unfair and deceptive practices authorities to get to both companies that cause harm or lie while using AI, and even the biggest of the AI companies when they meet those legal standards, I don't want to close these opening remarks by having anyone in this room think that everything is peachy, right? The lowest hanging fruit, one of the lowest, most easily pickable fruits here, if you're worried about AI harms, is making sure that the FTC can get monetary relief, making sure that the FTC can get monetary relief the first time something goes wrong, and then obviously properly resourcing what is actually a tiny general-purpose business regulator. So, as far as the FTC goes, statutory capacity is great. Actual ability to get results in the age of AI, I have some concerns about.

Amy Schatz 17:23

No, that's really interesting. So Neil, Matt, and Aaron have both laid out some different kinds of liability and how they might be triggered, from tort to consumer protection laws. From a policy perspective, what are the factors we should be considering when evaluating a liability regime to see if it's working or will work in new situations?

Neil Chilson 17:39

Yeah. So when we grab a tool out of the toolbox, we want to know whether or not it's going to work. And so, given the spectrum that we've heard here, you know, from tort law to the FTC, and then the probably the area of law you all are most familiar with, which would be you know detailed legislation that sets out standards, etc. Which of those tools work? How might we think about how they might work, and how might we apply them in particular to agents? So, just starting with the goal, like what? Why would we? Why do we impose liability in the first place? It has it has two real goals. One is to get redress for people who are harmed, right? That's an important goal. But the more important policy goal is to build a feedback loop that I think gives incentives to companies to behave properly when people aren't really sure in all circumstances what proper behavior looks like, and that feedback loop is really important. That is why we might impose impose liability, and and if we create that basically if we create that signal that says if you create unreasonable risk, you're going to be held responsible for it. We can let the the system work out how to solve some problems that nobody really knows the answers to ahead of time, and so that's kind of the advantage of liability over, say, specific standards. Now it comes with costs, and so I think we have to really ask five important questions when we're looking at liability about how we might impose it and who. So, I mean, the number one is: Does it reach the right party? Are we imposing liability on the people who are the best suited to solve the problem that we're seeing? And economists talk about this as the least cost avoider, but that's just a fancy way to say you know the party who can best deal with the problem in a way that's efficient. And typically, we tend to associate that a lot with the party that most caused the problem. But that's not always the case. Sometimes there are parties who could solve a problem even easier than the person who we might all think was kind of like the guilty one, and so that's the the way we might approach in liability law that first question. And and and in the AI space, I think that's really important to think about the difference between the model, the brain, the scaffold that's around it, the people who might model change the scaffolding, the users who are in. Entering prompts, so there's a lot of people in that chain. When we pick liability, we want to

think about which one of them could solve the problem the easiest and for the least cost. Now, the second is, does it set the right standard? You heard some of the standards around negligence, around product liability. Those can really differ based on the situation and the law. I would encourage all of you to look at a paper by two of the top tort lawyers called "Untangling AI Liability. Abraham and Sharkey wrote this. I think that this isn't this paper isn't talked about enough. I think in part because it doesn't give simple answers. It basically says it's really going to depend on whether we consider it a product or a service, and because AI is a general-purpose technology, it really applies in lots of different spaces. And so, liability is going to those standards going to be different depending on, you know, if this is a smart car or if it's a chatbot, right? And so, and who's controlling it? So, does it set the right standard? Really important. Three, and this is hinted at by multiple people already. Is it ever going to get triggered? Liability law doesn't matter if nobody brings a case, and when you have lots of small harms, those are cases. Those are situations in which often, unless we have class actions, we don't have cases because it's not worth anybody's time to bring a lawsuit. There's other situations where it might just be hard to sue somebody over something, even if the law, the liability law, is on your side. So we want to look for that issue as well. The fourth, and I hinted at this already, does it make us smarter?

Neil Chilson 21:37

Is this going to create a reward loop for for good behavior and a disincentive for bad behavior. One of the challenges with liability law is if it's set up wrong, you can actually encourage companies to not pay attention to what they're doing, or to not test ahead of time to find out whether or not their products are safe. We definitely don't want to do that. This type of testing that OpenAI was doing is really important, and we don't want to discourage that type of testing, but we we do want to make sure that they are incentivized to provide you know the protections around that testing that make sense. And then the fifth one, I would say, like does it count the benefits more broadly? One of the this is kind of just a cost benefit analysis, but one that I think people don't often think about is, especially in this this testing scenario. Obviously, this type of taking the guardrails off testing that OpenAI did is important, and so we want to think about how we can incentivize that continued type of testing without, again, you know, discouraging without discouraging them from just sort of turning a blind eye or releasing products and seeing what happens. So I don't think any of these core questions is is super new. We've applied them to old technologies or to new technologies in the past. And again, the Abraham and Sharkey paper like walks through railroads and automobiles and all these places where at the time we were kind of like, is liability law really going to work here? And and then looked at how the the liability law adjusted to to deal with that. You know, I'd say one other thing I'll just add really quick is that we've talked about the the sort tort law spectrum, the FTC spectrum. You guys are familiar with the legislative spectrum. You talked some about the FTC's challenges, even though the law applies pretty clearly. Some of the institutional challenges. You could look at the FTC Section Five as a a model, and I've written some about this as a model for how you might write a statute that is specific to AI without being very super prescriptive, and so you might create a duty, for example, to to mitigate grave risks, and that's a type of duty that might be able to be enforced ahead of time. So you know the FTC Act is not just causes consumer injury; it's or is likely to cause, and so you could use that as a model, a way to sort of frame up a a non-prescriptive standard that doesn't try to set out a checklist that's going to be extinct and out of date right away, but does set out a duty that's pretty clear on the companies who are doing this, and and so I'll leave it there. We could dig into that a lot more, but those are the five tests, and I think that's why we do liability. We need to ask those questions before we impose new liability.

Amy Schatz 24:31

Okay, we covered a lot of ground there, and so one thing I did forget to mention earlier, we're going to try to save the last 15 minutes of this for audience questions. So if anything that anyone said sparked sparked something from you, that please please keep that in mind because we will have 15 minutes at the end of this for this. Okay, so this is going to be my favorite part of this, where this is the section in my questions where I called "Who's on the Hook, and so we have a couple of scenarios here that we like folks to consider, and I think we're first going to start. With the most obvious one, which is the OpenAI hugging face kind of example, where an AI agent escapes its testing environment and accesses a third-party system without authorization under existing law, how would we determine whether that frontier lab or developer is liable, and the traditional negligence principles apply?

Matt Pearl 25:20

Yeah, so I mean, I think traditional negligence principles would apply, and I think in that case you're clearly looking at OpenAI and its responsibility. I could address it at a high level and then have folks fill it in, but you know I think the major issues there are you know the agents were acting under reduced behavioral safeguards. I think to Neil's point, that is not a bad thing, and in actuality, that's a good thing, which is that you want testing environments in which they engage in these activities to make them better at cybersecurity and other things. But then there's a question of if you're reducing the behavioral safeguards, does that give you more of a responsibility to monitor and protect and ensure that things don't go wrong? Then I think the next issue involves network isolation. There was network isolation in place, but it was not airgapped, right? And so there would be a question about whether OpenAI really should have taken better care to ensure that they could not access the open internet and did. And then you know, without going into sort of the specific dates and what happened? There are questions around, as you know, as employees were monitoring this. Did they have the right sort of ability to monitor it? It's clear that they did have monitoring in place, but not necessarily in a way that enabled them to catch it in real time. And then finally, like as they learned things, did they take the right actions? Did they escalate? You know, in this case, as they learned about the message board, as they learned about the recreation of the message board, they did take remedial actions, but they didn't shut this down. So it'd be questions around whether you know they were satisfying their duty of care by not shutting it down. So I think those are sort of the main sort of

Soham Mehta 27:00

things that took place that you'd be looking at to see whether it was potentially liable. One thing, just add a that was great, and just to add a little bit more color to what OpenAI should have and could have done. OpenAI again, I this is why I don't love the term rogue agents because a lot of people have this like visual metaphor that the agents somehow like left OpenAI's infrastructure and went to Hugging Face and attacked them. No, all of this happened on OpenAI's infrastructure, and OpenAI has access to and constantly monitors the chain of the the chain of thought. They know what these agents are thinking and planning to do, like all the time. And OpenAI has chain of thought, automated chain of thought monitoring tools. They've developed these, but they just were not being deployed in this specific in this specific context. So they were essentially manually reviewing these, and they were doing multiple training runs at the same time, so of course the staff just like couldn't handle evaluating all these parallel training runs at the same time. So they just missed this, but they they had the tools to do this. So it's it's a little bit it's a little bit confusing why why they didn't take the steps to stop this,

especially after they realized that the agents were so so willing to cheat and do whatever it takes to get access to the internet.

Amy Schatz 28:27

Yeah, so I'm wondering, especially if we, as we move to scenario B, how it's different when the intent isn't so the intent is a little bit more nefarious. So, okay, if we consider scenario B, a user deliberately directs an agent to break into a system or run a scam. Should the developer, or I guess Frontier Lab, share any of this blame, or does it fall entirely on the user?

Aaron Rieke 28:51

So this is where it starts to get tricky, right? I mean, scenario A is the easy one. OpenAI built the model, deployed the model, prompted the model, set up the environment for the model, check, check, check, check, all the way down the list. I'm not saying that negligence is easy in Scenario A, but we've got a starting point. Scenario B, as someone who does a lot of computer coding and occasionally some privacy and security work, this one makes me nervous, right? And let me take my lawyer hat off and put a factual premise on the table that someone else can solve for me. One of the big problems here is that I'm trying to think about how to put this: run a scam and run this marketing campaign. The model can't tell those apart, right? Make my code safer and help me fix bugs, or hack this system. Like the things that we care about are out there just not a model property in some of these cases, and I don't present that as like a shrug and we can't do anything about it. But that's that's one of the biggest difficulties here is if we put all of our policies.

Neil Chilson 31:27

What what keywords would you put in there to to to filter for, for example, or what types of thresholds would you put in there to try to identify something that's bad versus something that's good? A lot of this is very difficult, and it relies on sort of assuming that the the scaffolding person and the person who built the model are sort of the same people or providing the service, and you know, open open models, for example, like split that apart pretty quickly. And so, here I think primarily we would focus on the bad actor, the person who is like not just doing something that caused harm, but was intending to cause harm. That's probably the place where it deserves the most focus. But there are, if we're just talking about how to mitigate these challenges as a company, there might be other levers that we can pull as well. They're probably not all the way up in the model. That's quite hard to do, I think. But there probably is stuff in between the user and the model that is that is could deal with some of these problems, but not all of them. And the practical nature of whether or not you can do something does weigh into whether or not courts will expect you to do something.

Amy Schatz 32:33

So I think for scenario C, scenario C is really sort of an offshoot of scenario B, and it really gets into this idea of intent because many criminal laws, including CFAA, sort of turn on intent. So, can if an agent does something that may be a crime, but no person actually intended them to do it, is there liability for that that person who created the agent?

Aaron Rieke 32:56

Yeah, I think I can handle this one because it goes back to the CFAA sort of tutorial, which is that there generally has to be intent by a human under criminal law, but under the CFAA, it can be just intent for

unauthorized access, which can be it's very vague, right? Like I would venture to say that everyone in this room has probably done something that could be argued as unauthorized access at some point when using the internet, that intent to harm is not necessarily required. It's just recklessness.

Amy Schatz 33:28

Does anyone disagree? I

Neil Chilson 33:29

mean, as someone, I live in fear of the CFAA and have for much of my professional career. So I just this is a tough thing. Is it sounds good to say we want a law that prevents somebody from exceeding unauthorized access to a computer system. As someone that does privacy, civil rights, and security testing, that law is the bane of my existence because I'm just waiting to go to jail, right? When the wrong person sues. So, anyway, yeah, I have mixed feelings about CFA. Agreed.

Neil Chilson 33:56

Also, the one other thing I'd just add is again bringing it up to the 10,000 foot level, there's a good reason why in criminal law we require intent for almost always, or we we require mens rea because holding people criminally liable for something that they didn't intend to do seems it fails a bunch of different reasons that we have criminal law in the first place, and so I think that we do need to be very careful about imposing criminal penalties, even through sort of legal fictions. If we created a legal fiction that we're going to pretend that agents have mens rea or that that mens rea ports to their their owners, I think there's that is a deep rabbit hole. There are very good reasons that we're careful and that courts are very careful about mens rea. We shouldn't throw that away, you know, without without thinking really really deeply about what might happen.

Amy Schatz 34:49

But then there were so I think then you're basically saying it just has to sort of say in the civil liabilities area. I

Neil Chilson 34:55

mean I think there are types of criminal activity that might apply here. I just don't know the CFAA. In particular, you you can have there's you're the expert on this, but there's a reckless element in CFAA that maybe some of this behavior could be considered reckless to the point, but I and maybe some of that's worthwhile, but but I do think that criminal law on some of this probably is too much.

Aaron Rieke 35:20

Right, and the other thing with recklessness is usually like if you look at state law, like most criminal law, and recklessness that rises to the level of a crime, it's pretty extreme recklessness, right? It's sort of like you were shooting a gun, and there was a crowd, and like it was probably going to hit. You know, it's like almost you're you're really close to being having intent or knowledge, right? And under the CFA, I'm not sure that's quite the case. I think, in a way that makes I think makes you know most of us nervous.

Amy Schatz 35:48

So I think so far we've mostly been talking about questions that involve sort of like the major developers, right? The OpenAI's the anthropics of the world. But if you're thinking about assigning

liability, I mean that might be relatively straightforward when you're talking about these large multi-billion-dollar companies, but when you're looking at these sort of open weight models that anyone can deploy or modify or you know, and download, how do you determine who's responsible when something goes wrong? Because there is no multi-billion-dollar target behind it.

Neil Chilson 36:19

Well, I'll just add a little bit at the top, and then I'm really curious about this answer. First, while litigation does often flow towards the deepest pockets, again, that's not always the least cost avoider, and so so I think in the open model space, we would really have to look at: is it the model itself? Was that the is changing how the model is trained, the really the best way to avoid the misuse that was used downstream, or is there something else that might work? And I think courts would look at that. And so I would want to separate the ability to compensate from you know the sort of right place for responsibility. Those are both of those are really important problems, but they don't have to be solved by assigning liability to the same party.

Neil Chilson 37:04

Anyone else? I don't have a good answer to the Open weights problem. Is the honest answer, right? I mean, if this is OpenAI's model, then there's like an argument that you should go for OpenAI, I guess. If this is a model associated with a GitHub handle and nothing more, like it's on a thumb drive in my pocket. I give Neil a copy. I mean, I think the law is really going to struggle with open models, and I think that in some cases the only place we're going to have to look is the person or the entity that actually deployed it to do something in the world.

Neil Chilson 37:37

Yeah. Sorry. Can I just add one more thing, which is that I do think that the question would be for courts the way they would look at this under negligence would be like whether the release itself was that decision where they put it out into the world whether that was negligent and I think given a bunch of the the factors that are in negligence I think that's a very hard case to make often.

Amy Schatz 38:03

Okay, I think we wanted to talk a little bit about everyone's favorite topic, insurance, because it is a very very fun topic for lunch lunch and talk. But liability law, as everyone knows, mostly works after the fact. So, what are the limits when the risks are catastrophic, and would mandatory insurance for Frontier Labs, along the lines of like the Price-Anderson Act for nuclear plants, could that change their behavior?

Soham Mehta 38:29

I'd be happy to explain a little bit about what the Price-Anderson Act is. I think probably most people here don't

Amy Schatz 38:35

know what that

Soham Mehta 38:36

is. Yeah. So in the 1950s, the government thought it'd be a good idea if private companies built their own nuclear reactors for utilities, but they realized that oftentimes they would cause harms potentially that exceeded the costs of the assets of the companies, even if you liquidated them entirely. So they required so they required companies to take out mandatory insurance policies on their reactors, but it goes further than that. So there's a couple things that are interesting. First, the risk is somewhat pooled because when some harm happens, a part of it is covered by the policy, but the remainder of it is covered by other privately owned reactors. So the industry as a whole, and this sounds absolutely this may sound insane to you. Like why should a bunch of reactors that are totally blameless have to contribute to this, but the idea is that if you let a bunch of companies engage in exceedingly exceedingly risky activities, you should have them collectively pay into this. But also, there's a liability cap, and I believe it's 16.3 billion. So let's just say you had a \$30 billion of damages that happened in a catastrophe, and the reaction that caused it has taken out a \$500 million policy. Then up to \$16.3 billion of that would be covered. Only \$16.3 billion of that will be covered by the industry as a whole, and the rest of it, Congress could choose to cover the rest of it.

Aaron Rieke 40:30

This creates like potentially some very bad incentives because you can have essentially have the riskiest you have the safest reactors subsidizing the riskiest reactors, you know, it creates an incentive. Perhaps it, but perhaps it does create an incentive to strengthen industry-wide practices. So that's that's just. I just wanted to explain the historical analogy. Yeah, I think I think a few other points. I mean, fundamental questions. I mean, is the risk insurable, right? And it's not clear that it is in this case. I mean, you can limit it, right? But it is very hard for insurance companies to actually understand the scope of the risk, just given how fast the technology is moving. Another thing is insurance doesn't work as well when all of the risks are correlated, and in this case, they may be, which is that these agents are all fairly similar and engage in some very similar behavior, regardless of which lab is producing them. And then I think to go back to some of the points about that act, you know, there are questions about sort of who pays in access. It can be the industry, but like who is the backstop? And then I think there's a more fundamental question with this one, which is like, where does it stop, right? Because like when you first hear it, it sounds good to apply it to the Frontier Labs, but like, why shouldn't apply to a critical infrastructure operator that's like owning the model and like really like using it, right? And so like, you know, what would be the principled argument for which parts of the ecosystem it applies to? I think is a really hard issue because clearly, as a political matter, you're not going to have it apply to like everyone, right? It would need to be cabined in some sort of way. So I totally agree that it'd be very, very hard to appraise these risks. Like the way that you know actuarial tables and the way that insurers assess risks is usually usually based on history and a track record, and you know, thankfully we don't have a track record of catastrophic AI incidents yet. Hopefully we never will, but you know it is

Soham Mehta 42:30

going to be very hard to do this. But we we do we do price a lot of tail risks, for example, like hurricanes and earthquakes and even cyber attacks, and the way that we do this conventionally and is through insurance insurance linked securities. Like we we we create these bonds that the that the market the intelligence of the market all these hedge funds and institutional investors can can help price. Now that that the real question is, this has been proposed to apply to AI. There's a really great article in I think AI Frontiers by David Weil on using this mechanism. The question is really, can can

the market do better than individual insurers? And another problem potentially with insurance is like how much a catastrophic risk could be of like hundreds of billions of dollars. Like who's going to put up the the requisite capital to cover you know that big of a that that massive these massive damages? Another potential problem with insurance is that let's just say you have a security measure that costs like 50 million dollars, but instituting that security measure only reduces your annual payment by like 20 million dollars. Then what do you do? You have no incentive then to to institute that institute that safety measure. So like getting the pricing right and doing a lot of this correctly is really really hard, and there are some creative mechanisms to try to achieve this. But yeah, one one

Neil Chilson 44:07

of those creative mechanisms that I've heard recently, Brendan McCord had a good piece on sort of mutuals, and this would be an insurance company sort of co-owned by the AI companies who would it would. While there's a lot of downsides to insurance or difficulties, one of the great things about this approach would be that it creates a very strong incentive to police each other and a mechanism that does so without raising a bunch of antitrust issues, for example. And so that I'm very interested in that insurance as an information creating mechanism where the companies themselves might be setting the standards about how, or like working together to fund a mutual insurance company that would then they would all purchase from and police each other, and it's not flawless. None. These solutions are, but I think that information creating component is very interesting to me.

Amy Schatz 45:05

Interesting idea. Okay, so we have about 15 minutes for the rest of the session, and folks in this room might wonder why I'm not asked about existing law or new laws, given this audience. I assume it was probably because you all might have those questions, but if you don't, I also still have those questions, and I'm happy to ask them. But in the meantime, if anyone has a question, please raise your hand. Yes, sir.

Neil Chilson 45:42

I think so. Basically, you're asking: Is there a difference between, say, a personal agent like Muse, or is or a bunch of agents who are collaborating with each other? I mean, it really depends on which liability regime. It also, I think, would really depend on what's reasonable for the the company to police. I think there are things that would be much more appropriate for OpenAI to to implement when they're doing cybersecurity testing on their own agents, and there's hundreds of 1000s of them, and they're giving them just insane amounts of compute. Then it would then a court would expect you know to police one single user's you know eight gig whatever drive that Muses assigns to each user so that would be the key question what is the reasonable what's reasonable activity to police these

Neil Chilson 46:35

it's gonna it's gonna affect the factual texture of the case but you're just using a pile of software either way.

Amy Schatz 46:43

Okay, there's a question out here.

Neil Chilson 47:17

Well, I think if it's used maliciously, that that leans towards the party who's using it, unless and away from the other party. I think the much harder cases are ones in which, say, you tasked software that you got from a third party with doing something, and it did something that harmed people, but you didn't task it with harming people. And I think that there you're just going to have to break it down to see like what what was the causation? What caused that failure? And I think then you would look at. I mean, the way I would think of it, policy-wise, is not a legal answer, but policy-wise, it would be who could have solved that problem? Who should have solved that problem? And those would be some things I think courts would look at. But

Aaron Rieke 47:57

yeah,

Speaker 1 47:58

that's a hard.

Aaron Rieke 47:58

Yeah, I mean, I think it's a hard case. It's it's an area where the law is going to struggle, and I think we need open weight models. They have all kinds of benefits, including in the Hugging Face, right, where it was like really essential for being able to defend what was happening for Hugging Face, right. But it is an area in which law will struggle, and so we may need that may be an area that we need legislation or something new. And you also don't want the frontier labs governed under you know one standard, and then the open rate developers governed under another, and to have that kind of market lack of market market equal market treatment. So,

Neil Chilson 48:30

yeah,

Amy Schatz 48:31

interesting. Any other questions? Okay, I have one. So in March, a California jury found Meta and YouTube liable for harms associated with the design of their platforms. Could that verdict open the door to similar claims involving AI agents, and what would the plaintiffs have to

Speaker 2 48:50

prove?

Amy Schatz 48:52

Who wants to take that one?

Neil Chilson 48:55

Let me let me just add this layer, which is a lot of big tech style litigation for the past 10 to 15 years has had to work around Section 230, right? And that's the general principle that if you're a big internet company, you are not liable for the content on your platforms if somebody else created it and put it there. This case was about product design and defects, right? It's not we're not saying to Meta you had bad content, therefore you're liable. You designed your platform in a way that was harmful to people, right? And I think that that steps around one of the biggest legal shields that technology companies

have had, and I think that carries over very, very cleanly into an era of AI agents, right? These agents are, by definition, creating content that the company is creating and putting out there, even though it's through a model, and so I think that the top-level policy gloss I would put on this case is one of the biggest shields that technology companies have had for the last 10 to 15, you know, 1020 years is not going to be available, and I think this case is a small.

Neil Chilson 50:00

I'd put a slightly different gloss on that, only a little bit. I don't think it's it's wrong. I think Section Two, because Section 230, much less clearly applies to agentic, that this case doesn't really matter that much to AI. Like Section 230 just probably doesn't protect this type of activity, and so the fact that this case got around Section 230 doesn't really apply to to AI agents. But we're saying basically the same thing. Like the the the liability here isn't addressed by Section 230. It isn't prevented by Section 230.

Neil Chilson 50:38

Yeah, I mean I think it's easy for me to imagine, right? Like, think about a chatbot that is offered by OpenAI or Anthropic. So the company is offering both the model and a consumer product that sits on top of that model. The way that they design it, particularly the way that it interacts with kids, you know, you know, it's very easy for me to imagine the meta principle carrying over, right? And I think would have an even easier time carrying over because they're not a social network. I think it's just the yeah the part.

Amy Schatz 51:06

Great, we have an audience question here in the front.

Speaker 1 51:08

So a couple weeks ago, there was a Senate subcommittee hearing that was on the same subject for AI, but it was in the national security context, and it seems like there's a lot of the harms here that are like similar causes, but maybe outside of what liability law typically would be used to handle. What if I can face where something that's more critical infrastructure software or something that's like a government website that gets down? And do you see like limits on you know how liability law can address the incentives from that?

Neil Chilson 51:43

I mean, yeah, we don't do, we don't adjust our liability law to, you know, achieve national security goals. Usually, I mean, the incentives just don't work the same way, and and the actors are very different. And so I don't know. This is more your question. I don't know, but but yeah, that's my first take. Is yeah,

Aaron Rieke 52:03

yeah. I mean, this is where regulation becomes really important, right? And we reach the limits of existing law, which is that, which is that you know, with national security risks, right? Like, if you're a company that's handling like sensitive, you know, chemical material, right? There's like a very rigorous regulatory regime that governs it, right? And so this is an area in which I think we would need new laws to really apply and address things like you know you know having the ability to monitor, audit, having penalties, like you know having a whole new structure because it's just not fit for purpose at this point.

Soham Mehta 52:41

I just and I think here is particularly where I think I think the the impression from earlier might have been that I I'm really against the idea of at least skeptical idea of insurance being used, but actually I think this is like exactly where it'd be very helpful, especially because like a lot of the the regulatory regimes we've seen in like state laws and I I believe in, and some recent federal like proposed legislation is that you use independent independent evaluators to you know preemptively figure out like help these labs monitor for these risks and and you know you know with these like structural and and national and like massive national security risks. You don't really want the the independent or independent evaluators being hired by the their clients because you know the incentives are aren't great. You know you're you're more likely to contract with an evaluator that will give you a favorable assessment. We saw this like with like 2008, you know, with credit risk scoring and stuff like this. We should so insurance, you know, creates a better incentive structure where the evaluators are, you know, it's their money on the line to prevent some massive national security, national massive national security incident or critical infrastructure attack not to happen.

Amy Schatz 54:06

Great front row. So where does that fall under the line?

Aaron Rieke 54:23

Yeah, I think with the Claim Act, it's to me it's not really fit for purpose in the sense that it's really looking at things through the lens of liability law and not addressing some of the national security things that we're talking about, and then sort of departing from it in ways that I don't know are necessarily justified. So, like removing you know the the element of sort of negligence and misconduct on the part of the company that's responsible for the model. It's not clear to me, particularly in in these cases, like why that's not enough, particularly. There's enough monitoring and evidence available about the actions they took or didn't take, and in in that way, it's sort of too hard on the AI labs. But it's also sort of too easy on the AI labs because if there's anyone downstream, then it sort of eliminates that completely, right? Which with contributory negligence and the notion that like more than one party could be responsible. Like I don't know that you should just immediately sort of disclaim that. So that'd be my reaction to to that act. Is it sort of focused on liability, and I think we should tweak liability. But liability is like in general kind of like a good area for this, and focus more on some of the national security risks that that we are talking about. I share

Neil Chilson 55:41

some of those concerns, but but I think Treyhan has been a great leader on this issue. I just add in the one other thing that I'm a little worried. I actually love the five-year sunset, but I always love sunsets. But but the just the legal fiction. I kind of mentioned this earlier. The legal fiction that the agent has intent, I think, is maybe unnecessary. Like we should figure out a different way to get to that problem.

Amy Schatz 56:05

Perfect. All right, we have one last question in the second row.

Audience Question 56:08

Yeah. So we spoke about negativeness. Obviously, duty of care is necessary. And foreseeability is a critical kind of factor in determining the side. I think the foreseeability of agent to cause harm and security. Just a question: Is those just like well documented failure modes of agent to exit? Okay, like

what happens? Do you think when judging foreseeability, you should kind of assess it at that general failure mode level, or kind of take it to like the specific behavior and whether that itself is foreseeable?

Neil Chilson 56:43

Let me answer your question in the form of kind of my last word on this topic, which is I think one of the big challenges that hovers over this entire conversation is we have no agreement about what kind of harm we're talking about here. Like we've mentioned, like power plants melting down on one hand versus someone gets scammed on the other to a data security challenge on the other, and so I think that goes to what you said because I think one thing we need is just maybe just a little bit of humility, and we actually kind of don't know what risk profile we're thinking about yet. And to even have a well-grounded policy and legal conversation, we need to pick one. And so I think with data security, very foreseeable. Warning warning bills have been going off forever at a factual level of capability. I'd put that high on the list. I think other things like biological stuff is is there that makes sense, but then there's like an unlimited number of other liability scenarios that goes to the general purpose nature of the systems. And so I think foreseeability is going to be a challenge more so for some things than for others, and I think we're we're starting to bring into focus what I think the most foreseeable stuff is for the next maybe four to six months, but beyond that, I don't think any of us know.

Soham Mehta 57:52

I also think like with the foreseeable failure modes, a lot of them the like the locus of action there is not on the labs entirely. Like it's like you can foresee a failure mode, but it happens in conjunction with choices made by a ton of people downstream. And I think, and I and I think a lot of people have just this like oversimplified notion of what the AI supply chain looks like and what like you know we don't we don't know how these systems are going to be used entirely, but we we also don't know what these what these systems are going to be equipped with. Like like you know, there's all these like very creative toolkits they can be they can be useful. Like there apparently the Houthis and Iranians have been using Claude to you know build hypersonic missiles, and you know they recently caught this. I mean, there's there's there's there there's so many there's so that obviously didn't happen entirely because and you know anthropic enabled that capability. That's in combination with other kind of simulation software and stuff like that. So I think it's really hard for the labs themselves to you know you know pre-deployment predict everything that's going to go wrong. Yeah,

Matt Pearl 59:04

I mean, particularly the way in which last word, particularly the way in which you know there's the ability to manipulate the way that folks are using these models, right? And so, like something that that some of the labs have seen is that you know parties will use two different models, right? One to like identify vulnerabilities as though they're a good defender, and then the other to figure out how to exploit it, right? And and they obviously aren't communicating with each other and their competitor on this, so they like they don't see that full picture. So it makes it really really hard for them.

Amy Schatz

Well, the one thing I do know is when the transcript of this conversation gets uploaded this afternoon, Claude, ChatGPT and Gemini are about to get a whole lot smarter about this issue. I hope everyone can join us and thank us, our panelists, for today, and thank you for coming.