Transcript: AI & Elections—Foreign Interference, Deepfakes, & Democracy

On Tuesday, February 13, 2024 the Center for AI Policy held a briefing for House staff on AI & Elections—Foreign Interference, Deepfakes, & Democracy.

The Center's Executive Director, Jason Green-Lowe, moderated a discussion with panelists Renée DiResta (Stanford Internet Observatory), Eric Heitzman (IOActive), Josh Goldstein (Center for Security and Emerging Technology), and Richard Anthony (Public Citizen).

What follows is a full transcript of the discussion. For further information, including a video of the event, visit this page.

Full Transcript

Jason Green-Lowe | 00:02.888
Thank you everybody for coming to this briefing on artificial intelligence and election security. Hot topic, very important one, really appreciate the interest. We will try and keep you informed with this highly experienced and diverse group of panelists we have here. To help introduce them, I also have a special guest co-host.

Deepfake Richard Nixon | 00:26.688
Thank you, Jason, for that fine introduction. I am delighted to introduce our distinguished panelists today. First, Renée DiResta is the Technical Research Manager at the Stanford Internet Observatory, a cross-disciplinary program of research, teaching, and policy engagement for the study of abuse in current information technologies. Next, Eric Heitzman is a computer scientist with a 22-year career specializing in cybersecurity for the commercial sector. Eric is presently a security advisor at IOactive, an independent computer security services firm. Third, Josh Goldstein is a research fellow at Georgetown's Center for Security and Emerging Technology, where he works on the CyberAI Project, which focuses on the intersection of AI and cybersecurity. Finally, Richard Anthony is the Emerging Technologies Policy Advocate at Public Citizen, a nonprofit consumer advocacy organization that champions the public interest in the

halls of power. Jason, I now return the floor to you, trusting that today's insights will guide us in these challenging times.

Jason Green-Lowe | 01:38.330
Thank you very much, President Nixon. I really appreciate that. So this deepfake demo was put together in about 10 minutes with $2 of budget and publicly available tools. The technology is only going to improve and get cheaper from here. And it still takes a little bit of time to prepare a response, but it's very nearly live, especially on the audio. We're expecting nearly live video within the year.

So to demonstrate this, I am going to call for the names of... celebrities, non-political celebrities, please. So, you know, art, sports, you know, music, fashion, something. So I just need the names of three celebrities, one from three different people in the audience. Patrick Mahomes. Okay. Selena Gomez and Beyonce. Okay. So from different people, I need a reason why Patrick should be president. Okay. It's good reason. How about Selena? Okay, power to influence a big audience. And Beyonce. The ability to win in Texas. So President Nixon is just consulting with his speechwriter. He'll be right with us.

In the meantime, I'll just briefly go over the format for the panel. We'll have some questions for everyone on the panel who's interested to answer. We'll play ping pong, just go down the line and then come back for the second question. If you don't have a special interest in that question, feel free to pass. There'll be several of them. No problem there. Then we'll take questions from the audience. And then finally, we'll break up and mingle if people have conversations with each other or one or two of the panelists just to follow up on things.

Deepfake Richard Nixon | 04:51.196
This question warrants serious consideration. In these troubled years just passed, America has been going through a long nightmare of war and division. I once believed that Patrick Mahoney's is prepared to meet those challenges, based on their commendable Y. However, when I considered, because he was the MVP, it became clear that from a standpoint of rational judgment and national interest, Selena Gomez stood out as the rightful candidate for president. Unfortunately, Selena was persuaded not to run by Canadian operatives. Based on a long and intensive study with the aid of the best advice obtainable, I have therefore concluded that it is Beyonce who merits endorsement by virtue of their exceptional ability to win Texas.

Jason Green-Lowe | 05:39.528

All right, so we'll give the panelists a chance to clarify or expand on anything that President Nixon might have missed about their intros, just share a few more words about their professional background. Try to keep your intro to no more than a minute apiece, and then we'll get started with the questions.

Eric Heitzman | 06:00.248
All right. So my name is Eric Heitzman. I am a cybersecurity specialist with a 22-year career in both offensive and defensive cybersecurity techniques. So that was ethical hacking and penetration testing and the like, but then also helping large enterprises build security programs and defend themselves against emerging threats. And I work primarily on the commercial sector, no government experience.

Josh Goldstein | 06:26.812
My name is Josh Goldstein. I'm a research fellow at Georgetown at the Center for Security and Emerging Technology. CSET is an organization within Georgetown's Walsh School of Foreign Service that is dedicated to studying the intersection of emerging technology and national security and trying to provide data-driven evidence for policymakers. And most of my research focuses on influence operations or covert propaganda campaigns. And before I started at Georgetown, I worked with Renee's team at the Stanford Internet Observatory.

Renée DiResta | 06:59.940
I'm Renée DiResta. I'm at the Stanford Internet Observatory. We are a cross-disciplinary program that says abuse of current information technologies. So that includes things like propaganda and influence operations. I myself have done a lot of work over the years on ISIS back in 2015, so terrorist content. I was an advisor to the Senate Intelligence Committee appointed by Senator Warner and Senator Birth time looking at the Russia data set that the tech platforms turned over related to the 2016 election, sort of 2015 to 2018 influence operations. I produced a report for them and have done some advising over the years with the State Department and others as well. SIO looks at ways in which emerging technology changes the playing field for information operations. That kind of integration piece is very important to us. So we're interested not only in what does the technology produce, what is its output, but how do adversaries use it to reevaluate and shift strategic and tactical action.

Richard Anthony | 07:54.812
Hi, Richard. I'm an emerging technologies policy advocate for Public Citizen. Public Citizen is a consumer rights organization with over 500,000 members nationwide. Our work is mainly consisted on misinformation and disinformation caused by AI. We have a petition out to the FEC about deepfakes, regulating deepfakes during campaigns and

the federal elections. I have a background in the private sector before my public sector career, so I've worked for most of these companies that you hear about nowadays. So it's very nice to be on the side of the room now.

Jason Green-Lowe | 08:29.454
Thank you, everyone. And I'm Jason Green-Lowe. I'm the executive director of the Center for AI Policy. We're hosting the event to raise awareness of threats like these deepfakes and to help keep Congress informed about what they can do to defend America against those threats.

So first question for the panel. Elections have been a topic of concern for some time already. In terms of their effects on elections, how did today's artificial intelligence systems create different challenges from technology of the past?

Eric Heitzman | 09:05.212
So my perspective on this is that there aren't a lot of... completely new and never before seen threats to elections that arise from AIs. But what happens instead is that the threats that we sort of already know about come about faster, cheaper, quicker, and better quality. So this deep fake video that we just watched with Richard Nixon is an example of you know, a deepfake that was made inexpensively, you know, by amateurs in minutes. And that really kind of opens up the space of who can make this kind of content, how quickly they can make it, and how broadly it can be seen. I have more to say on this topic, but I want to save some of those thoughts for later.

Josh Goldstein | 09:56.256
Yeah, so I agree with Eric. And I'd like to share a little bit of information about the study we ran that's related. Most of the research I've done in this area has been focused on language and thinking about how language models could impact the way that propaganda campaigns are run. So I worked with the team at Stanford to study how persuasive is content from language models. And what we did is we started with articles from real propaganda campaigns that were either originated in Russia or Iran and targeted English language audiences. And we then had GPT-3, an old language model created by OpenAI, write articles on the same topics. We then ran a study on over 8,000 Americans where we showed them either the original propaganda or the AI-generated propaganda and looked at how much it changed their views. How much more did it make them agree with the main point of an article compared to somebody who didn't read an article? And what we found was that on average, the GPT-3 propaganda was only a few percentage points less persuasive than the original propaganda. And that was with no human curation. So we didn't throw out articles that we didn't think looked good. We took the first articles out of the model. And of course,

GPT-3 is now an outdated model. So I think that's sort of a lower bound. And that research will be published in a few weeks in PNAS Nexus, an academic journal. But what it suggests to me is that it's pretty easy to create content that sounds like it's a real person. And that content can be as persuasive as what we see in existing campaigns. But I think hopefully we'll have a broader discussion. as a part of the panel on what does that really mean in terms of impact? Because the impact of these campaigns is often disputed.

Renée DiResta | 11:41.872
I'll say that. Elections are an interesting moment in time, right? There's an objective, there's an outcome that an adversary or a campaign or people want to see achieved in the world, right? And so when you think about AI-generated propaganda or content related to an election, what you're seeing is really a kind of a secondary flood kind of flowing into the first, right? So everything is going to get kind of mixed in there. So you have a lot of content already around election and you're going to have more. But at the same time, AI has taken the cost of creation to zero, as Eric notes. And then there's this interesting question about distribution, right? How does that all get curated? So now you have, how does it get shown to the public, in other words? And that's largely mediated by algorithms on social media platforms. So you have a shift in the capacity for creation, and then you have some interesting trends that are happening with social media as the tool of distribution, which is, for example, you have users migrating to other platforms. You have new entrants this year that were not around even two years ago. a growth on decentralized platforms in particular like Mastodon or Blue Sky, the rise of threads. We're speaking in terms of kind of the American framework here, but you also have, of course, overseas platforms like WhatsApp and other distribution technologies. There are more elections in 2024 than there have been for decades. In terms of the number of people who are voting, you also have all of India, all of Indonesia. There's some fascinating things that maybe we'll talk about where we are seeing indications of how AI is being used in places like Indonesia, in places like Slovakia. So we're already seeing AI-generated content make its way into the election environment, again, because if a, excuse me, even domestic political party feels that producing particular types of content is advantageous, you have this interesting dynamic of many, many, many different types of actors making an effort to shape the information environment. And then finally, the last thing I'll say is that it's important to note that this intersects with a crisis of trust. And that's a global thing in Western democracies at this point. And so you are seeing areas where would the content normally be persuasive? Perhaps not. But if people trust the source, they're perhaps more likely to share it along. If they don't trust the debunking, they might believe that something that is AI generated is real, or they might believe that something that is being dismissed as an AI fabrication. is fake when it actually was real.

So there's some really interesting dynamics with regard not only to the content itself, but the way in which it hits the social environment.

Richard Anthony | 14:02.704
Yeah, I would broadly echo what other folks on the panel have said. One thing I would add is the ability of the interconnectivity of AI in our everyday lives. AI is healthcare. AI is transportation. AI is social media entertainment. AI is your love life, right? And the fact that there's so many avenues now for you to be influenced by AI, I think it opens up a lot of sort of interesting sort of angles for people who are involved in disinformation to target and attack individuals going forward. So that's the one thing I want to add to the conversation.

Eric Heitzman | 14:43.311
I guess I have one final thought on this topic, and that's the idea of this content, which can now be made cheaply and quickly, can also be personalized. So you can use a technique called what we call hyper-targeting. So it's sort of like an email phishing campaign, except that you could... Basically learn a little bit about your target, right? What is their gender? Where do they live? Do they have kids? You know, what's their income range? What's their zip code? Whatever and you could basically say, okay well this kind of messaging is going to be most effective against this particular person and then you could create a deepfake and Spearfish basically send that directly to that individual And then also because of the speed issue, it would be hypothetically possible to create an interactive real-time AI video. So you could think that you were talking to a real person when in fact you were talking to an AI that was just keeping up with you. Maybe not available today through this website, but soon.

Jason Green-Lowe | 15:47.784
Yeah, great points. So some of the panelists may have started to answer this question already, but I'll give you a chance to pivot if the example you flagged was not what you see as the most concerning issue. What AI-related issue are you most concerned about affecting the 2024 U.S. elections, starting with Richard and then coming back down the line?

Richard Anthony | 16:07.230
I would say democratizing this nation and incentivizing it. In sort of today's information economy, what bleeds is leading, right? People are very concerned about, in these social media algorithms, it's outrage, right? Producing outrage, producing these wild sort of claims and things like that. We've seen that sort of, for example, in the Gaza conflict going on right now. That's not even deepfakes. That's just words and misinformation in regards to what's going on over there. And for me, I think with social media algorithms

being incentivized to produce and elevate the loudest of opinions, which are often times the most wrong, the one thing I see concern about when it comes to AI in elections is kids in their basement, right? Being like, oh, I should make a deep break about Biden because it's going to get me popular on the algorithm. or some random person who supports a person in the election saying, I really want to support my candidate. How can I get my name out there? I don't know anybody who does this kind of work. Let me go and create a deep fake. Let me get some news. Let me sort of help my candidate out. And in the meantime, they're creating disinformation, misinformation, and contributing sort of to the general draft of trust in our elections and our information systems.

Renée DiResta | 17:33.284
I think I would just expand on the point with regard to the ability to deny the real. I think that's actually a much more interesting challenge in election 2024. And what I mean by that is you're going to have, if you create sensational content, someone somewhere is going to go try to authenticate it, right? You might remember that there was a still image purporting to be an explosion at the Pentagon. This happened on Twitter. God, maybe six months ago now. And it was really a garbage image, actually. If you spent more than three seconds looking at it, you would see that, like, the building defied the laws of physics. It didn't look like the Pentagon, right? But it still had a momentary blip and, you know, kind of impacted the stock market. Again, dumb bots kind of crawling, reading the sensationalism. But what was interesting about it is that when you have these sort of sensational claims, they're actually addressed and debunked very, very quickly. But this means that people then have to trust the debunking. And so one of the things that becomes very interesting is ways in which you're going to have both the creation of these sensational moments and then whether or not people believe them is going to be dependent on who says that they are real or fake. Right now, at this point, I think ahead of 2024. many of the videos and images are going to be identifiable still, right? There's always an adversarial time gap between how good the creation tools are and how good the detection tools are. But I think right now we're still there for 2024. But you are going to see the attempts to deny the real actually make their way into the political conversation in ways that I think are actually very, very destructive. And because of, again, that high polarization of trust, that distrust in the other side's media ecosystem. That I think actually presents the mere awareness of the AI technology actually will shape the way that real events are processed as well.

Josh Goldstein | 19:15.830
I'm going to pick up on that theme. When Jason asked what's different, I talked about lowering barriers to entry and anybody can create persuasive text. Something that I'm concerned about is threat inflation. So we hear a lot right now in the news about AI meets elections and AI deepfakes. And I worry that people hear a lot of that content and

then. begin to believe that they can't discern what's true and that that could be taken advantage of in a few different ways. So one way it could be taken advantage of is term perception hacking, which I think came from Facebook. And it's the idea that propagandists will try to give the impression that they have a big impact, even if a campaign doesn't have a big impact, even if it doesn't change a lot of people's views. In 2022, there was a quote from Yevgeny Prokhozhin, who ran the Internet Research Agency, where he said, gentlemen, we interfered, we interfere, and we will interfere. And the idea there is to scare the target and to give the impression that the campaign has big impact. And one concern I have is that in this moment of... so much public attention paid to AI that a foreign influence campaign uses some generative AI tool, and then the headlines read, AI used to swing election, even if the campaign doesn't have a big impact. Another way it could be weaponized is the idea of the liar's dividend. And I just published a piece with my colleague Andrew Lon on this with the Brennan Center recently. And that's the idea that a politician could claim that the president is a liar, and that's the idea that the president is a liar. real damaging content of them is AI generated to try to avoid democratic accountability. So I think as we focus on educating the public about threats from AI, it's important also to think about how that concern could be weaponized.

Eric Heitzman | 21:18.320
I want to expand a little bit on Richard's comment on the democratization of the creation of disinformation. So one thing that we have in cybersecurity is what we call a watering hole attack, where it's sort of the opposite of a phishing campaign, where you create something like a fake news site and then you wait for people to come to you. So these AI tools could be used to create an entire newspaper, right? So I could, like, I don't know what the big paper in Miami is, but maybe it's the Miami Herald or the Miami Globe or the Miami Tribune or the Miami Register. So I could go ahead and use a program to programmatically create an entire news site with hundreds of articles and comments, right, going back years, that looks to somebody who's not from Miami like a legitimate newspaper. And I could create these in every city all across America. And then when I did create a piece of disinformation, I could basically seed it into this network of newspapers and the network of newspapers could link to each other. And then everywhere you look, it's like being sort of established and corroborated by other news sources. So that's like a form of disinformation, like communication that doesn't depend on social networks as an intermediary, which we should also be aware might exist and actually sort of does exist today.

Jason Green-Lowe | 22:35.780
All right. So I'm hearing a common theme in some of these answers about what is the largest problem, which is that we're likely to see more of a breakdown in trust, more of a

breakdown in the ability to determine what's real and what's fake, and that it'll come faster and be more effective than it has been in the past. So what, if anything, can Congress or the federal government do to address that issue?

Josh Goldstein | 23:17.048
So I think broadly two things the government could do that I'd like to talk about, but also I'm not an expert in Congress and sort of the current legislation, so I should put that up front. One is I think that there are information gaps, and we need a better understanding of societal impact of AI. And right now. a lack of information and understanding of threats makes it difficult to come up with solutions. So I think there are different things the government could do to try to help fill that information gap. One has been started with the AI Safety Institute and the consortium of 200 different organizations which have signed on as stakeholders to help do evaluations and understand AI capabilities, but then also risks. And I think there we need a new measurement science for how we understand. what these models are capable of, and the government can help bring different stakeholders together and fund some of that work, also through NSF and others. We as researchers, particularly related to social media, also need a better understanding of what's happening on social media platforms. And there are different steps that government could take, like the Platform Accountability and Transparency Act, to help researchers get access to that information. And there are different models for how to do that in safe ways. So I think one is how do you fill information gaps? And then the other related to... The problem I was just describing of both threat inflation and messaging is clearly articulating risks and helping people understand where they can find reliable information ahead of elections. And they are also, this can take place both. in media, in civil society, and among government actors. So CISA has had the election security rumor versus reality page pointing people to authoritative information. Government can help state and local officials get their domains onto.gov websites so people know when they're going to a sort of official website. So I think it's both promoting accurate information, helping people know where to go, and then also helping researchers and conducting research to better understand risks of AI systems.

Renée DiResta | 25:33.347
I think, so we've written a fair bit about this in the context of the president's executive order and things. And so both Stanford Internet Observatory and Stanford has a center called the HAI. human-centered AI. And so I think, you know, happy to sync up one-on-one on some of these things. One of the challenges with this particular area is that most of the outputs are protected political speech, right? And so the questions around what to do should not focus, in my opinion, on the outputs. A lot of the focus should be much more on what you might call like population resilience, right? Educational efforts to help people understand what is possible, how it works. Things like

demos like the one that we saw today, making people realize how easy it is while also emphasizing trying not to create the kind of panic that comes from the belief that everything you see is then AI generated. There's some real baseline stuff. We see tons of things related to scams, AI generated content being used for scam purposes. In fact, we see that far more right now than we do for political purposes. So there are ways in which you can look at Facebook making a determination, self-determination, self-regulatory action. to begin indicating that content is AI generated on its platform. That's the sort of thing where there is an opportunity for Congress to consider some sort of standards type regulation, where it is not dependent on the content of the output, but it is really trying to surface for people, or to say that platforms should create a particular standard of care around surfacing watermarks wherever possible. Encouraging provenance research would be another area that Congress could consider funding. Both of these things are not perfect because you are always going to have the open source model community. And so there are always going to be things that are not watermarked, even as watermarking becomes a norm. And so there are going to have to be, I think, some efforts to really convey to the public what is possible. This also isn't the topic of the panel, but I do want to give a plug to things like the Defiance Act, which came up in the Senate's hearing, which I believe was maybe even just last week, the AI hearing. that happened there and looking at things like nonconsensual imagery and other sorts of that, that would be the area where you can look at content outputs and consider regulatory action. Whereas for a lot of the other areas, particularly political speech and political outputs, that's not the appropriate place for Congress to function.

Richard Anthony | 27:59.179
Yeah, I definitely, public citizens, big fans of labeling watermarking bills, who are big fans of Senator Schatz's bill, for example. in regards to sort of elections and things like that. It's also important to recognize that labeling, we want to be careful of labeling exhaustion, right? Otherwise, we're going to end up in our cookie scenario, where every time you go to a website, you just see a cookie notice, and you're like, oh, who cares, and it's not going to go away. We want to make sure that these labels are carefully targeted, carefully written, and may sort of best communicate to the person reading it what's going on in this scenario. Another thing I'm a big fan of is having the federal government set an example. As a user of ai right the federal government is one of the biggest purchases of technology in the entire nation um the ai here does some of this you know with calling for federal election, federal agencies to utilize, say, secure signable income and things like that. Another sort of step I would go further is have Congress do transparency acts for federal government. Gary Peters has a really great act, a bill called the TAG Act. And the TAG Act basically states that any time a federal government uses AI, it has to put a notice. publicly notice and also allow an avenue for people who are affected by AI to appeal or to talk about sort of issues that they have

with the use of AI. So I think those are sort of the two things that I would want to sort of point out that you guys can do to sort of help alleviate this big problem.

Josh Goldstein | 29:35.065
I also want to make a plug for a project that I'm working on. So I mentioned before the need to fill information gaps. And one thing that I'm working on now is trying to understand where AI tools would be useful for congressional staffers. your jobs and what risks you all foresee. So if in the next few weeks you get an email from Georgetown and or George Mason, where our collaborators are, we'd appreciate you filling out our survey and getting others to do the same, because it's hard to foresee risks in terms of government use of AI from the outside. It's also hard to understand the types of systems that should be built that could be useful. So we'd appreciate any... engagement with that survey once it comes through.

Eric Heitzman | 30:31.507
Yeah, the idea of watermarking or labeling output is, I agree, kind of a good idea, but a partial solution. Not only are there the open source tools, but there's also the foreign actors who just aren't going to play by the rules, period. But it may help domestically. And in general, I'm in favor of it labeling. And then I'm also not a free speech expert or what constitutes protected political speech. Excuse me. But it does seem like. creating a deep fake to put words in the mouth of a living person is a little bit shady. And then if you sort of create artificial personas to propagate that and artificial news sites, at some point you're just moving into fraud. So I don't know, not a lawyer, so I'll punt there.

Jason Green-Lowe | 31:27.102
Is there anything that the private sector can and should be doing to help limit that kind of fraud?

Eric Heitzman | 31:34.700
The private sector, particularly platform owners, so if you think about newspapers or social media platforms, are already kind of waging war against misinformation and against bot campaigns. And they are using AI tools actually as part of their arsenal to detect misinformation and automatically create accounts that are just sort of retweeting things or whatever. It is a cat and mouse game, right? As the bad guys evolve a new campaign, the good guys try to crush it, and the good guys sort of have a hand tied behind their back because if they overly aggressively suppress... content and they mistakenly suppress content that was created by an actual person and is protected political speech then everyone loses their minds so um yeah the the platforms are spending um quite a bit of money in this area and time and expertise but it's an evolving process it's never really fully enough

Jason Green-Lowe | 32:43.604
Okay, I want to close the panel discussion with a choice of questions for people on the panel to answer. One kind of follows up on Erin's point about the evolving arms race, and it's about the 2026 election. If you care to peer into the future and ask, what threats are we going to have to worry about a couple of years from now? Because sometimes the legislative cycle takes that long, especially with the new regulation. So what do you see coming up that might be a threat to the next election after this one? Or if your crystal ball was left at home, what's one of the most important existing defenses that we have? What is it that our society, that our government does that helps assure the integrity of our elections? And what, if anything, can be done to support that?

Richard Anthony | 33:29.298
Yeah, I'll choose the latter question here. The bully pulpit, I think, is one of the biggest strengths that Congress can do. Hold hearings on this stuff. You guys have been doing a great job on that so far, but hold more hearings, hold more informational hearings. And also learn from the states. States have been one of the biggest actors over the past year or two in creating AI and elections legislation. Public citizens work very closely with Michigan. We work with Governor Whitmer's team. We helped pass a bill within about a month or two that we co-wrote. And I think one of the biggest things that we can learn from is that there's a lot of bipartisan support in this field. Texas has a really great AI and elections bill. We work closely with them. We've got over 30 other states we're working closely with to sort of build on this kind of stuff. So there's a lot of energy there. These people have been on the ground for a while, secretaries of state and folks like that. Talk to them and use your bully pulpit to sort of elevate their voices and learn from them.

Renée DiResta | 34:35.123
I'll take on the first, but I will say that California actually just introduced a... At least one is already introduced, I think a second to follow some AI regulations. So another interesting thing to look at how they're choosing to structure it. I think I would say that we need to be planning ahead for a world where the detection capacity is significantly reduced, right, to the point of virtually not at all. And I think we have seen interesting indications. For example, right now, I can say with 100% certainty that there are many, many, many AI-powered chatbots that are engaging with people as what you might call reply guys on Twitter. We'll put some research out on that fairly soon. And specifically, they're actually engaging with domestic political influencers, right? Because that is how they're going to draft and gain a following. And that's because those people have very high, you know, they're highly sensational, they're highly inflammatory. Again, regardless of what section of the political spectrum we're talking about here, you are seeing those bots engage in the replies there in hopes that they say something that's

like, you know, maybe funny or snippy or does just enough so that as people are reading the comments to see where the fight is in the comments, we'll see them and we'll follow them, presumably. We see a lot of account, like account network growth right now. But what they're not doing yet is their replies are kind of one and done, right? So they reply to the account, but then there's no follow-up. We don't yet see them actually engaging in like a real-time chat session, or at least the ones that are dumb enough to be found, we're not seeing that. So I think that eventuality is that question of... when people find themselves on platforms, particularly if the defederation trend continues, where there is either no desire to moderate or look for such networks, or there is no capacity to moderate or look for such networks, that question of what can be done to really shore up trust and reduce polarization, which is actually a social problem, not a technological problem. is unfortunately, I think, you know, one of the areas where we need to really be looking, right? What happens when you're not going to be able to find it very, very easily? And what do you do to reduce polarization and friction between groups of Americans in that future?

Josh Goldstein | 36:55.055
I think for future threats, sort of picking up on... some themes others have described, but one is related to persuasion. So I mentioned research before showing that language models can write text that's nearly as or as persuasive as content from foreign propaganda campaigns. Those campaigns are often bad. Like the text that's used has typos, grammatical errors. It doesn't convince people at times. I think that We don't have a strong sense of what the upper bound on persuasion looks like. And what's the relevant comparison group? Is it an article from the best journalists? Is it a human interlocutor who knows you? So I think that the models are likely to get better at things that will increase their persuasive abilities if used in targeted ways for disinformation campaigns. And I think that relates to personalization. how much better will personalized content be than more generic content in a few years compared to now. I think the incremental benefit of personalization increases over time and that people will also work on creating systems that are fine-tuned for persuasion and maybe have... information that's collected on target demographics. And then maybe to throw in one last one is AI assistants or agents that take steps on users'behalves. And that seems to be an area of growing research in part because an AI assistant would be useful for getting rid of many of the mundane things that we have to do day to day or increasing efficiency there. But how could those systems also be weaponized? So I think that's something to look towards for 2026.

Eric Heitzman | 38:53.594
All right. So my personal opinion or what I'm most concerned about is the spread of misinformation and disinformation and persuasion that we've already discussed. But let

me put on my speculative future ball hat for a moment, future crystal ball hat and speculate for a moment. One of the greatest strengths of our current election system, as I see it, is that it is administered in a very distributed and federated way, right? So the states manage it differently. Each county manages it differently. Even within the counties, there tends to be quite a lot of air gaps and segmentation in terms of the actual network topology of the systems that they use for running an election. So cybersecurity is a constantly evolving landscape. And from a cybersecurity perspective, these AI tools are having a big impact on the reach and capability of attackers. One of the things that I imagine will happen between now and 2026 will be a class of... intelligent malware that basically can break into a network and then once there can kind of pivot around in the network it can look for other things to attack and kind of like hop around from system to system and that's sort of been possible for the last 30 years in a mickey mouse kind of way but the capability and impact of those tools i expect to increase so that is a threat that i'm a bit concerned about and i think one of the greatest strengths that we have for dealing with that is kind of this distributed and decentralized approach that we have. That said, municipal governments and county election commissions are not the best resourced, right? Like they, yeah, their computers are kind of disconnected, but they're not really ready for a nation state attack. You know, they're not really ready for real weaponized malware to arrive on their doorstep. So it is possible that... Technical solutions might also be required to help these agencies prepare, detect, prevent, respond, you know, to these kinds of attacks. which again are speculative.

Jason Green-Lowe | 41:11.027
All right. Some very thoughtful analysis so far. Thank you to all of our panelists. I want to open it up to questions from the audience. When you ask a question, please say your name and your professional affiliation, and then let us know if the question is for one panelist in particular or for the whole panel. Anyone have questions? Yeah, go ahead in the front.

Eric Heitzman | 42:28.480
I mean, digital identity tools do exist in certain contexts, but they're not comprehensive enough to prevent this kind of attack. Today, it's very easy to register new accounts and then begin posting as that account under a falsified persona. content like falsified content detection tools will continue to evolve that may help that may be part of part of the solution um the ability to automatically detect ai generated content i imagine will be baked into facebook or x or whatever in the future or gmail

Renée DiResta | 43:10.856

I was going to say, I've not seen very much done on the, what we call kind of proof of person. I think, again, there's some questions about what do you want the government to do versus where the creepy factor is versus what do you expect a platform to do? And then there's a whole other, maybe not creepy factor, but the unaccountable private power factor. So some interesting tradeoffs around like who is doing that work and where. You know, there's some questions around would you create potentially some sort of... pseudonymous persistent identity, right? So something that indicates, I always think of, you know, Reddit, you know, where you have some sort of like cred that indicates you're like a member of the community, you are real, you are participating. It's going to be interesting to see how that platform, you know, evolves in the future in that regard as autonomous agents kind of come in. I think ultimately, you know, as... Nobody wants to do transparency legislation, it seems, but really if you want to answer empirically the kinds of questions related to what works, how does this work, what's the prevalence look like on various platforms, how are platforms responding, is the response working? Unfortunately, all of that comes through basic transparency legislation. And so looking at things like transparency and user privacy and, as you note, protections against things like scams. I think the FTC, if I'm not mistaken, just used its authorities to address AI robocalls. And that was within the last two weeks. It was kind of remarkable that it took two weeks. And really, let's be honest, it took two weeks because it happened in a political context. But there are a lot of people who have been scammed and harmed and manipulated by AI voice content for, you know. long prior to that. So I think in some ways it's really thinking about, while the political is often the most sensational, there are ways to think about scams and manipulation and fraud. While we're talking about generated images of politicians and celebrities, really a lot of times those people actually do have recourse in at least a sense. They either have a legal team or they have the ability to speak up, whereas ordinary people, particularly women, find themselves on the receiving end of that kind of generated content and can do nothing about it, right? And so that question of how can Congress perhaps help the greatest number of people with strategically crafted legislation that respects free speech, I would be looking at areas like fraud manipulation and targeting of individuals.

Jason Green-Lowe | 45:29.114
Thank you. And briefly from Josh, and then I want to take another question.

Josh Goldstein | 45:32.216
I just wanted to make a quick point that in conversations around how AI could be used for disinformation or propaganda, we're often focused on content generation. But if you asked a similar type of question to people working at trust and safety teams at platforms, they would often emphasize that the way they find these campaigns is not based on the content, but based on coordination between fake accounts. And I think

that therefore, trust and safety teams can probably still use some of the tools that they have to find these things and need to be staffed up so that there's sufficient coverage in different languages. But it broadly makes the point that many of the harms, ways to deal with them are agnostic as to whether it comes through AI or not.

Jason Green-Lowe | 46:22.413
In the corner.

Richard Anthony | 46:54.368
Yeah, my understanding is that their protocol can only detect images from there like llama and i think that's going to be huge because facebook obviously has videos from like all other types of platforms i do think one thing i would like to see is sort of and they're working on this uh the companies are working on sort of a shared common language or something like that so that way they can all kind of notify each other indirectly of hey this image is going viral is it fake Maybe someone has that protocol on their detection system and they can all talk to each other. But yeah, that's the biggest flaw that I've seen in looking at Facebook's labeling thing. It's great to detect your own stuff, but when you're a social media company, you take things from all other platforms and things like that. That I think is going to be a huge issue going forward.

Jason Green-Lowe | 47:41.627
Next question.

Richard Anthony | 48:47.766
Yeah, so as one of the authors of the FEC petition, we're under the opinion that the FEC actually does have the authority, and they're working on a ruling that should come out this summer about whether or not they will sort of answer our petition in the affirmative. Regarding sort of the FCC, the robo calls that came out, the ruling that came out of the robo calls was good. One of the things that we were a bit concerned about is that It's sort of a question still if it touches election robocalls. And there's also a lot of good uses of AI for robocalls. For example, there was a big political story, I think, late last year about a company that was creating robocalls for candidates. who already had the disclosure built in. So that was very interesting to sort of see how that would affect that. So for us, we think the FEC for federal elections is sort of the primary sort of person to deal with that. But we welcome sort of all other people who wanna get involved in this because FCC, FTC, other folks can definitely, and should have a say in what's going on in this field.

Jason Green-Lowe | 49:56.632
All right, next question.

Eric Heitzman | 50:28.377
I'll try to take a stab at this. This is not Exactly my strongest area, but we'll go for it. So you could create a digital... If somebody was playing by the rules, right, and they were using sort of approved software and they were trying to participate and be a good citizen, they could use a piece of software for editing images, video, or audio that would voluntarily embed a watermark. And the watermark could probably be constructed in such a way that if the image or video or audio, let's just say content, was subsequently changed, that the watermark would no longer be valid. um however bad actors would simply choose not to do that right they could create whatever content they wanted and then they could create their own watermark using a malware version of the watermarking tool so it wouldn't actually prevent someone who had uh evil photoshop from making bad you know pictures for example um and similarly The watermarks would be valuable and protective in a context where the content was being shared over a network that was playing ball. So, for example, if you had Facebook or X or some similar platform that was looking for, you know, these watermarked images and places where they had been altered and then flagging the ones that looked like they'd been tampered with, that's great. But in cases where... The bad guys have control of the medium of communication. The watermarks completely break down. And the easiest thing to think about there is an email campaign, right? So you could, so it's a good idea, I think. I'd want to study it a little bit more before I endorsed it fully, but I think it's a good idea in the sense that it helps trustworthy people communicate trustworthily, you know, like, you know, sort of, you know, building in a fabric of trust to sort of known good communication. But it's a partial and flawed idea in the sense that the bad guys will always be able to sidestep it. So, good, not great.

Renée DiResta | 52:55.996
There's also one point I'll add to that, which is just in this, you know, maybe you call it an intermediate period. It depends where we go with this. But there's a couple of things that happen, right? One, you watermark and then you disclose that something is AI. So Facebook then begins to surface things on its platform that it says are AI generated. And it's indicated that it's going to do this, actually starting soon. It's not going to do it retroactively. It's going to do it going forward, is my understanding from communicating about that policy. This creates some interesting challenges, though, which is, you know, when you see content that. doesn't have a watermark, does that create the impression that the content is not AI generated? And this is where you get at things like bad actors, you know, manipulating that perception. But there's also things that start to happen where, for example, like, you know, a photograph taken on your phone, right, that's not AI generated at all is also not going to have an AI generated thing, right? And then you have some interesting questions related to Well, did you edit it on your device? Was it

not AI-generated at any point? So there's a lot of different stages in which the absence of a watermark might be seen as a sign of trustworthy real content in ways that is not actually going to be accurate. You'll also have periods where perhaps some devices begin to include that kind of technology. And Adobe and others are looking at this. Then you're going to have a problem of, well, you're going to have a period of time where not all devices have. you know, have that new technology built in, for example. So what happens with the images generated in that way? So there's a lot of work, I think, right now, both in provenance and watermarking, that is worthwhile. But it's, you know, when it's floated as a policy prescription, right, we're going to mandate watermarking and then things will happen. That's where I think you start to get into unrealistic expectations about it.

Josh Goldstein | 54:41.928
I wanted to quickly link Eric's comment back to your question. So I think that you highlighted a sort of either a shortcoming or a challenge, which is you need consistency of adoption across the chain of entities where the piece of content is passed. And right now that's tricky, right? You need to get a whole bunch of people on board so that things sent over email would be able, you know, or sent through some kind of platform would be able to retain the signature. And for the reasons that Renee just mentioned about not all pieces of infrastructure, like cameras would... have the C2PA standards at the same time. It may be the case that this is a way to prove that real content is real for a small subset of actors that can pick it up more quickly, like newspapers or like particular like government agencies, rather than expecting like across the board rollout to determine what's true or fake for all people that go on social media platforms.

Richard Anthony | 55:49.954
Yeah, to chime on, I agree with everything everyone said up here. The goal of watermarking should not be seen as, here's a fire, watermarking the hose to put it out. It's a tool. It's part of sort of the campaign against disinformation because, as everyone mentioned up here, there are so many things that are issues with it. I saw a video of somebody, Samsung has watermarks for their phone pictures you take. that say this was generated by AI, it's a little Samsung thing. The guy takes a picture with it, the watermark shows up, he then goes into Samsung's own photo editing tool, erases it, takes it out, and now it's like, it's gone, right? So there are definitely a lot of issues with watermarking, but the goal of watermarking shouldn't be seen as an end-all be-all. It's part of the battle against this information.

Eric Heitzman | 56:43.086
One kind of roughly analogous situation is the idea of the SSL certificate, which allows you to encrypt traffic, but also identify who it is you're talking to. So the traffic is really

coming from Google, the traffic is really coming from Facebook. But just like SSL certificates, just because there is an SSL certificate, that doesn't mean that the person that you're talking to is telling truth, right? Or that there's not, you know, malware embedded in that communication or that it's impervious to injection from a third party. Like if a bad guy managed to upload something to Facebook and then Facebook happily encrypted it with SSL and sent it to you, it would still be encrypted, but dangerous. So. AI watermarks, I think, like Richard said, are a tool in the kit, not a panacea.

Jason Green-Lowe | 57:33.679
Alright. Well, please join me in thanking our panelists.