The First Amendment and Artificial Intelligence

Thursday, November 5, 2026

Visit our media library at constitutioncenter.org/medialibrary to see a list of resources mentioned throughout this program, listen to the episode, and more.

[00:00:00.1] Julie Silverbrook: From the National Constitution Center in Philadelphia, this is *We the People*.

[music]

[00:00:08.1] Julie Silverbrook: I'm Julie Silverbrook, Chief Content and Learning Officer. The National Constitution Center is a nonpartisan nonprofit chartered by Congress to increase awareness and understanding of the Constitution among the American people. How can constitutional values help guide emerging technologies? What will the future of free expression look like in the digital age? In this episode, technology and constitutional scholar Jonathan Zittrain and Distinguished Professor of Law at Rutgers Law School Ellen Goodman explore the intersection of the First Amendment and artificial intelligence. This program is presented in partnership with the John S. and James L. Knight Foundation as part of First Amendment Day, a new annual celebration that invites visitors of all ages to explore the five freedoms of the First Amendment through conversation, creativity, history, and hands-on experiences.

[applause]

[00:01:01.8] Julie Silverbrook: Good afternoon, everyone, and happy First Amendment Day. So glad to see all of you here. The civic holiday you didn't know was a civic holiday, but it is one, and we're so pleased to have you here. And we're so pleased to have Jonathan and Ellen, and we're so grateful to the Knight Foundation for their support of First Amendment Day and this conversation. I want to get started by talking about the triad and the three camps that are emerging around AI. You call them the accelerationists, the safetyists, and the skeptics. Explain who falls into each of those three categories, what they're most concerned about, and where questions about free expression fit into those competing visions for AI's future.

[00:01:42.7] Jonathan Zittrain: Well, thank you very much, Julie. Great to be on stage with you, Ellen, and hello to everybody here and to the thousand folks online, many of whom are people.

[laughter]

[00:01:54.8] Julie Silverbrook: Not all of whom, though.

[00:01:56.1] Jonathan Zittrain: Great to be here, especially for... Did you say this was the first First Amendment Day?

[00:02:00.0] Julie Silverbrook: We're saying inaugural because first First Amendment sounds redundant.

[00:02:03.0] Jonathan Zittrain: Yeah, no, I was thinking the second First Amendment Day is gonna be weird, and then the first Second Amendment... Anyway. So for those of us who love the Ninth Amendment, we have some waiting to do. [laughter] But I have been hungry to make sense of what maybe we all feel is a little bit of a drip, drip, drip of news about things that approximate AI, and to make sense of it rather than how they once called history "one damn thing after another," and how to make it not be that. And one way to do that is to make sense of some of the consistent arguments that we've seen in the public sphere now, at least since generative AI first became widely known in the fall of 2022, and the different cultures around those arguments. So we have here a standard issue triangle, and here's how I've been thinking about it. As you said, Julie, on one corner are the accelerationists.

[00:03:07.3] Jonathan Zittrain: And you can kind of place them in San Francisco as their home base next to Starfleet Academy. And they are terrifically, generally excited about the prospects that these cluster of technologies we call AI, what they can offer. And the sooner we bring them to apotheosis, the better. I think their bumper sticker, if they had one, would be like, "Why wait to cure cancer? Let's do it now." And we just need a few more clusters and it will be done. And that's your standard issue accelerationist. And then across the Bay Bridge to the east, you have the safetyists. And they're like, "Oh yeah, we've got clusters, but it's a different kind of cluster." It's a... [laughter] yeah, well, you can... I realize it's First Amendment Day, but there was that *Schenck v. United States*, the draft case that... Anyway.

[00:04:06.6] Jonathan Zittrain: So the safetyists... Inside joke for the law professors there. For the safetyists, they have long been, even prior, for many of them, prior to LLMs coming into view, they have been really concerned about the prospect of intelligence that might either surpass, outnumber, or both, human intelligence. And asking the question, why would we expect that it would do whatever we want? Is simply training it on *Star Trek*, where Commander Data was a hero... Am I right?... Is that enough to keep it aligned? And I say east across the bridge because that's kind of the Bay Area, and Berkeley is the headquarters. There's a place there called Constellation, which is a kind of coworking space for a lot of safetyist organizations that are trying to figure out either how to slow things down or, depending on who you ask, put a cork in it entirely, or how to innovate in ways of making sure that the things are either not acting out on their own or not acting at the behest of Dr. Evil... Perfectly aligned to Dr. Evil, but misaligned maybe to the rest of us.

[00:05:20.2] Jonathan Zittrain: They actually share in common a sense that this AI will be transformative, something that the third group... If we're still going with our Bay Area map, they are actually in the water... The skeptics, who bring to bear a quite venerable tradition of science and technology studies and a kind of gimlet-eyed look at all sorts of tech, for which AI is, I think, for many of them, one in a series, that says, "We've seen this movie before. People get excited and hyped up about stuff, and we are often bringing our own hopes and expectations to it." Pareidolia is real. Being able to see a face in a piece of toast is a gift, but also sometimes an error. And to anthropomorphize this tech, as it so invites us to do, they would say is a mistake. And moreover, it pays forward so many, and makes more efficient so many, of the flaws and biases and encroachment upon the already marginalized that we have seen.

[00:06:31.9] Jonathan Zittrain: And so if we were just to organize them along a little graph here, I'd say you have the accelerationists generally thinking it's gonna be good, with a kind of rump faction saying, "No, no, it might kill us all. That's why we need to get there first." [laughter] So there's kind of a *Dr. Strangelove* group within that. And then on the bad side are the safetyists and the skeptics, who turn out, like Maoists and Leninists, not to like each other culturally, because the safetyists think that the skeptics aren't worrying about what's important... They're shuffling deck chairs while it might doom us all... And the skeptics think the safetyists are spinning science fiction tales while day by day AI is working its way into our daily life. There are all sorts of tweaks we could make to that graph. There are people who call themselves normalists. Everybody hates their label.

[00:07:22.7] Jonathan Zittrain: Please take a ticket if you want to complain. They all want to call themselves realists, but that would make a tough triangle if it was just realists, realists, and realists. So that's my way of trying to say, when you see events in the news, it's often easy, especially with the voices that the media selects or that the Twitter algorithms select, they will often fall surprisingly neatly into those categories.

[00:07:50.3] Julie Silverbrook: Ellen, I want to bring you in on this and ask you what you make of that framework, whether you think it captures the debate as you see it, or if there are dimensions that were missing in the triangle.

[00:08:01.8] Ellen Goodman: How can I object to that scheme? So I'm gonna answer your question, but let me maybe level set a little bit about the First Amendment and its application here. So the speech clause of the First Amendment, as most of you know, says Congress shall make no law abridging the freedom of speech and of the press. Okay, so in the context of AI, what might we be concerned about in the speech context and what might Congress or the states try to regulate? So I think as a first order concern, we're concerned about losing cognitive autonomy, about deskilling, about offloading our thinking process to machines. And the First Amendment, at root, those are its words, but at root, its conception is that we need to be able... We free people in a republic need to be able to think what we will and express what we might, and that machines are interfering with that. Okay, so how might they interfere with that?

[00:09:08.7] Ellen Goodman: I can just list a few of the concerns, and you should feel free to add to them. But we know, in addition to this loss of cognitive sovereignty, which is sort of abstract, although any teacher will tell you that they're seeing the effects of it very clearly, we might be concerned about chatbots and what they say to us and sort of AI psychosis and harm to children, scams, highly persuasive, personalized bot speech that misinforms or deceives us, a general sort of loss of grip on what's real, sort of epistemic confusion about what we're seeing. So those are categories of concerns that we're seeing states are beginning to try to regulate. And these are all... And I think we'll get into it... There's a First Amendment shadow over some of these efforts. Okay, so now taking Jonathan's scheme, I think if you're an accelerationist, you might say, "Okay, these are worrisome, but technology always poses these problems."

[00:10:24.9] Ellen Goodman: The radio did, the telegraph did. People were concerned about the telegraph, the radio, TV, and humanity adjusts and will adjust, and these are not sort of... These concerns are not different in kind from those we've had before. Safetyists, I think, think that... And I would include myself in that camp. Think that this is a different order of challenge to expressive freedom and to other kinds of freedom. And so therefore we need to meet the moment with interventions that are different from the ones we've had in the past. And this poses a real challenge both to the way we think about the First Amendment and the way we think about freedom of speech.

[00:11:13.5] Julie Silverbrook: Let's talk a little bit about many of the challenges, but one of them is that AI is advancing so much faster than our institutions, our law, journalism, are accustomed to moving. How much of the concerns people have about the technology have to do with the fact that it adapts so much more quickly than we as a society are able to catch up to it?

[00:11:39.9] Ellen Goodman: So I think it's both/and. It's really the delta between how fast the tech is moving and how slowly our institutions are moving. And so I just would, if you have been following the AI news in the last couple of days, you've seen the chief scientific officer at OpenAI, Jakub Pachocki, himself has just written something called "The Alien Mind," essentially saying they in the labs don't understand this technology and it's really out of control. So that's the technology moving very fast. And what he's saying is, "Please, please, you people out there, stop us." There needs to be international coordination and we need to be slowed down. And that's the institutional piece, is that we don't have those institutions that are able to act in that way. And in fact, I think at a time when we need institutions to move very quickly, they seem to be frozen like deer in the headlight.

[00:12:36.5] Julie Silverbrook: I think they're terrified. Actually, a lot of the institutions would be skeptics, not even safetyists. You make of that the same way?

[00:12:44.5] Jonathan Zittrain: Well, I think it's kind of funny to imagine OpenAI just at headquarters handing out pitchforks to people, [chuckle] being like, "Please, poke us." And that is some of the puzzlement of our time. I think on this transformative-boring axis, there's a

roughly proxied one for slow and fast. And there might be two kinds of threats we've already talked about, one of which is the classic safetyist worry that's almost like a lab leak or a *Jurassic Park* kind of thing, that they're testing their latest model called the "break out of a sandbox" model, and they put it in a sandbox. They have no idea why it left, and who could have predicted? And that is a kind of pure threat from pure tech that isn't widely distributed that a lot of the safetyists worry about. One of the things you may hear about is something called RSI, recursive self-improvement, that they have the models train their successors to be a little better, and then without taking a lunch break, the successors train their successors.

[00:13:57.9] Jonathan Zittrain: And in the parlance, instead of it being "boom," it's "foom," which is how quick it goes. So that's right to your fast or slow question. But I want to separate that from a really good list that Ellen just sort of bookmarked for us all that is, I think, largely in skeptic territory, which are all of the more quotidian ways that this tech can hurt us. And it may be a little slower as you have a deployment gap. It has to get out of that beaker intentionally. This is not a breakout now. This is, "Have we got a great model for you all?" And to intermediaries, "Have we got a great model for you to offer?" If you are a medical practice, it's gonna help your patients because it's gonna call them and have a conversation with them that you never had time to do to remind them of their upcoming appointment or to take their pills and to answer their questions. And you repeat that across a bunch of domains.

[00:14:53.4] Jonathan Zittrain: And that might represent some of the cognitive offloading, some of the other problems you were mentioning, for which that might naturally be slower and, depending on the domain, could allow time for the law to catch up. And so we have to walk and chew gum. It's just, again, I'm sort of taxonomizing here, thinking about some of the outlying pure tech risk that might be like a lightning bolt from stuff that hasn't been distributed yet versus stuff that is out and about. And name your school district: are they embracing this because they want to make sure the kids know how to use this stuff because it's everywhere, or are they saying, "Absolutely not, we don't want it to corrupt our kids"? Talk about a lack of agreement on what lies ahead.

[00:15:37.6] Julie Silverbrook: Let's take a step back and talk about the different types of AI. So I think a lot of people are familiar with chatbots, right? That's probably how most people are interacting with AI. But then there is actually AI that can go, speak, communicate, act on your behalf, in some instances without human intervention. So, Jonathan, let's bring out the dogs to talk about the different types of AI.

[00:15:58.8] Jonathan Zittrain: Very good. So there's all sorts of types of AI, some of which are just madly categorizing things. You give it a bunch of examples of how things are categorized. One of my favorite examples is Duchenne versus non-Duchenne smiles, which, as psychologists will tell you, are real smiles versus, "I'm not so happy, but I gotta smile" smiles. And you train them on enough, and pretty soon you've got something you could run against your Facebook or Twitter feed and be like, "They're not happy," "They're happy," that kind of thing.

That's so-called discriminative AI in a technical term. There's generative AI, which can generate new stuff, pictures, vocabulary and such. And then what you were talking about, Julie, is so-called, and I hate this adjective, but here we are, agentic AI, which just is the noun agent made into an adjective by adding "ic" to it. And we have another chart. This is the last one, I promise.

[00:16:54.8] Julie Silverbrook: There's pictures on this one, though, so.

[00:16:56.6] Jonathan Zittrain: So on the one axis, we have how powerful it is, which amounts ultimately to its risk potential. On the other axis, for an agent that is capable of first being given a general goal and using its "reasoning power"... The skeptics would be saying big quotes here. It makes up a plan and is off to the races, maybe just saying, "I'll be back at midnight, don't wait up." The other is that it can act outside the chat box. It can go into the world and just do things, often because it can pull the same levers humans can. It can click on things, it can look at websites, et cetera. And the third is you can set it and forget it. Some of these things may go on for a while. So for agentic AI, if it's really domesticated and not that powerful, it's like a Pomeranian. And what's not to like about Pomeranians? I was recently bitten by a Pomeranian-like dog, [laughter] but we have... We don't teach it at most law schools, but there's a lot of dog law out there.

[00:18:00.6] Jonathan Zittrain: There's a lot of cases, like slip and fall, car accident, dog bite. And we know how to handle this. The law is on top of it. And you register your dogs, and there's an animal commission and that kind of thing, and there's antibiotics for about three weeks. If it's domesticated, it means there's some human tasker for it or a lab behind it that actually powers it. And if you pull the plug on OpenAI, all the Pomeranians go to sleep that are OpenAI Pomeranians. It's a service, not a product. If you get a little more powerful but still domesticated, now you're into your pit bull or your Doberman kind of thing, and a dog that somebody might employ to go out and have... I know there might be Doberman lovers here, but you gotta admit they have strong jaws. And in the law, we have ways of dealing with those too. There are breed-specific laws that say if you've got a pit bull or pit bull mutt, you gotta do extra things in order to cabin the hazards the way you should.

[00:19:04.1] Jonathan Zittrain: Once you get to feral, you get to creatures that are not tied back either in legal or ethical responsibility or in technical terms, which can vary, to an obvious or any owner, whether somebody prompting them or a lab that powers them. And in our canine taxonomy, if it is feral but not that powerful, it's like a baby fox. It's out there doing its thing, who knows why. It's like some of those websites where you can see tumbleweeds running through them. There could be agents of that sort as well, possibly running on open-weight models that aren't tied to any lab. And if it's powerful and feral, now you're in the zone of the wolf. And wolves are really interesting. They can be single or they can be in packs. And many of the legal interventions and remedies and frameworks that we and others have been talking about to deal with rogue AI or AI that causes harm against expectation is really dealing with pit bulls.

[00:20:09.4] Jonathan Zittrain: Some of you may have heard of the more recent OpenAI Hugging Face... I can't believe I just said those words together, I'd like to send them back to 2022... But incident where the OpenAI agents broke out of their box and visited Hugging Face and broke in and all sorts of things happened. That is a pit bull story. That is a domesticated but powerful agent that has a lot of people rethinking just how much they are a safetyist. It's not even yet a wolf. And dealing with the wolf, including in what remedies we might do, prophylactic ones included, that wouldn't run afoul of the First Amendment, to me, is one of the big challenges of our time. And it includes possibly building a wolf sanctuary, a place where we can actually generate them, invite them in, try to understand them, see how they communicate with each other and how much that is legible to us. Unlike *Jurassic Park*, they're gonna be at the door one way or the other, even if they don't escape.

[00:21:12.2] Julie Silverbrook: The wolf is scary. Some of the other dogs are scary too. The Pomeranian could even be dangerous. So, Ellen, there are already attempts to regulate. Before we get to sort of the regulatory environment, I want to talk about First Amendment and the First Amendment interests. We're in a really interesting time where human speech and human thought are being combined with machine-created words and images and art. And then there are instances where the AI is creating its own art and images with not a whole lot of human intervention. Where are the First Amendment interests here, and whose First Amendment interests are involved? We both have the people who are writing the prompts of the AI, but then we actually have the engineers who are creating the AI. Can you walk us through just the sort of First Amendment landscape in the world of AI?

[00:22:07.6] Ellen Goodman: Yes. And you use the term, Julie, "First Amendment interests," which I think is really useful because there's First Amendment law, which I'll get to, and then there are First Amendment values. And sometimes there's a divergence between them. So, for example, First Amendment law doesn't have much to say about how do we promote an environment which the Knight Foundation cares so much about, a robust environment of arts and culture and journalism. First Amendment law just deals with government abridgment of speech. So I think there are a lot of issues here about First Amendment interests. And if we have time, we can talk about some of the copyright battles, some of the concern that machine-generated speech, AI-generated speech and content is displacing human generativity and journalism. Okay, but those are interests. On First Amendment law, I think the question is...

[00:23:06.8] Ellen Goodman: And it really goes to whether we're talking about wolves or Pomeranians and how strong and short that tether is between the machine and the human being... I think we don't yet know the answer to the question of is or are LLM or AI outputs covered by the First Amendment. We don't have any court holding on that or clear answer. What we do know from our First Amendment precedents, and they kind of cut in both directions, so I think courts would be free, judges would be free to go in either of these directions. On the one hand, from what I just quoted to you about the First Amendment, it protects speech, not speakers. Right? So you could imagine a court saying, "What we're concerned about are outputs. We don't

care how long that tether is." And what we don't want is government coming in and telling chatbots what they can say, what they can't say, whether they should be right-coded or left-coded.

[00:24:11.3] Ellen Goodman: In that sense, we would be in sort of familiar territory, and I guess it would be kind of an accelerationist perspective that this is nothing new. On the other hand, we also have law that is very much concerned, is very interested in the fact that speech is tethered to human beings. So even if we think about *Citizens United*, which protects the speech of corporations, well, corporations are associations of human beings. And so you could imagine a situation where if we're talking about agentic AI and what the agents are going around saying is so far from what any human being told them to say or even knows that they're saying, at that point, the arguments that we are no longer in First Amendment territory would be quite strong, I think.

[00:25:01.7] Julie Silverbrook: Jonathan, do you want to weigh in at all?

[00:25:04.3] Jonathan Zittrain: Just one question maybe to ask ourselves through Ellen's shepherding is how settled is this? How much can we turn to prior precedents? Like, well, maybe a printing press doesn't have First Amendment rights, but somebody using a printing press might. And if there were a law that said your printing press is capped if you belong to this political party at X RPM and it's double that for another, or you can't use it or something, there would, I think, be a very strong, to put it mildly, First Amendment claim that a new technological affordance that before anybody had it, nobody could have expected to use it, you are now entitled to use to go ahead and get your message out. And in many ways, large language models could be exactly that. Imagine commissioning... I don't think you have to imagine too much. A large language model that just... I don't know if you've seen the...

[00:26:05.8] Jonathan Zittrain: There's a cartoon called xkcd, and one of the panels of it from a long time ago is, "Honey, why don't you come to bed?" And you see the person's on a computer and says, "I can't, this is important." And the person in bed says, "What's the deal? Why do you need to do it?" And they say, "Someone is wrong on the internet." So suppose you believe that when someone is wrong on the internet, you need to burn the midnight oil, but you also get tired, and the price of freedom is eternal vigilance. So you have your bots go online and keep hitting back all the ping-pong balls and making all your arguments on your behalf. Would it be okay if a state, if the government tried to say, "No, we're gonna treat bot traffic differently," even if it is trying to be your extended reach the way the printing press was? I don't pretend to have an answer to that, but that's an example at least of a human interest behind the bot.

[00:27:04.6] Jonathan Zittrain: Rather than asking if you've got a fox or a wolf that's tapping the mic and says, "I have something to say," and nobody wants to lay claim to it, is it okay to give First Amendment protections there? The best case I can make for that is, depending on how they're configured, you could see an incredibly salutary use for large language model-based

systems to represent the view that didn't find its way into the room and has X amount of credibility, reasoning, empirical validity... You name whatever the other parameters are. That's trying to use it to actually enhance the kinds of debates that Ellen was talking about as First Amendment values, for which trying to restrict that would suddenly feel not so great.

[00:27:51.6] Julie Silverbrook: So let's talk about that. What happens to the analysis as AI becomes more autonomous? We're in territory where not much human intervention needs to happen for AI to start being generative. Let's talk about the ways in which... I think what you just said, Jonathan, is actually really interesting. It can add value to the marketplace of ideas, but there may not be value for the machine from an expressive standpoint. It's more what we're receiving from the machine. How does the analysis change in that context?

[00:28:33.7] Ellen Goodman: It depends, right? Because we have cases going in both directions, that we're really just concerned about listeners' interests and speech. I think I'll just draw everyone's attention to a case that the Supreme Court had a couple of years ago. It was called the *Moody* case, and it was about social media. And the question there was whether or not social media companies have First Amendment rights to do content moderation. And the court held that they did. And this is the analogy game, right, that they are just like newspaper editors, that they have editorial policies that they propagate through their algorithmic feeds, and that that's protected against state intervention that tries to direct their moderation choices. But there was a very interesting reservation in that case, which wasn't presented, which was this question: well, what about if the algorithms are not really subject to human moderation? At that point, would the analogy work? And they reserved on that. And I think that's the question of the sort of autonomous agents. I just... One other thing I would say about that is the fight will be about whether or not they really are ever fully autonomous, right? Because what we're asking the labs to do is to rein them in, right, is to bring them more to the Pomeranian side. As long as we're asking the labs to do that, aren't they then under some human direction?

[00:30:00.4] Julie Silverbrook: I want to talk a little bit about the printing press, journalism. And Ellen, this will be for you. You've written about how technology is changing our information environment. We're all aware of this, from deepfakes to creating art that's not actually... Well, is it art? That's an interesting question we could explore... But not created by the human hand. What are the opportunities and challenges in journalism? Because in every space that AI exists, right, it both threatens to undermine aspects of journalist... Journalistic practice, but it also offers, through large language models, opportunities for analysis, looking at evidence that we just as human beings, it would have taken us decades to do, and now we can do it in minutes or hours.

[00:30:47.8] Ellen Goodman: Okay, so in all of the AI debates, when you have the accelerationists pitted against the safetyists, the accelerationists will say, "I hear you, safetyists, and the best way to mitigate against the risks that you've identified is to have more AI." Right? You're worried about cyber threats? We need more AI to defeat cyber threats. It's the same, I

think, in journalism, where you can say first social media and now AI are undermining the foundations, the business model for journalism. And so what's to be done? We still need journalism to do what journalism has always done, which is to ferret out and report the truth. But so that's the... How do we combat that risk? Well, agentic AI, large language models, but specifically agentic AI. If journalists have agents running around, right, they are much more empowered. And also if citizens have agents working for them in the contest between sense and nonsense, right, having your own agents as sense-making bots or agents is going to empower you.

[00:32:03.7] Ellen Goodman: So there's that debate. I think now just putting on my Luddite hat for a second, I think there is no substitute for investing in people to do the work of journalism. And I just want to shout out an idea that's been floated by Steve Waldman, who's done a lot to try to... It's called Rebuilding Local News... To try to find various policy solutions to the problem of the decimation of journalism. And he's proposed that we ought to have sort of a token tax on data centers, that data centers, among the many other things that we ought to squeeze out of the data center build, is new revenue for journalism. And I think you and John Sands, Julie, have just written this really interesting op-ed about supporting institutions that do our sense-making, and I think journalism would be one of those.

[00:32:57.1] Julie Silverbrook: So let's talk about this. The AI developers, the labs are saying, "Stop, regulate us," right? In some instances, "Regulate us." When does that regulation start to potentially encroach inappropriately and unlawfully on potential First Amendment rights?

[00:33:16.6] Jonathan Zittrain: Well, I feel like Ellen already put a finger on one of the clearer ways that that could happen, which is we see the social function of LLM-based systems as being a kind of truth machine, an oracle that people turn to. I'd be curious to know within this audience where people are along that journey. If you haven't checked in since GPT-3, you might be figuring, "I ask it a question and it just hallucinates a bunch of junk. How does anybody use this stuff?" Then there are people that are like, the most idle question, you take a photo of a clump of leaves, you're like, "I think that's something in that," and you're like, "What are these leaves?" And one of the things about these systems are they will treat with the deepest gravity any question you give them. [laughter] You're giving them a scan, it's like, "Here's the biopsy. What does it mean, doc?" And it's like very grave. And then you're like, "Here's a pile of leaves. What do you see? Read them for me."

[00:34:19.9] Jonathan Zittrain: And that can be very powerful, especially if there is post-GPT-3 accuracy to it. And the labs at the moment are working on making those be as accurate as they can while disclaiming any truth. "This is for entertainment purposes only. Your scan is clear." [chuckle] And so that's very entertaining. And I think how to think about evaluating in an ongoing way these tools when people consult them for truth and they may fail weirdly, either in ways that no one anticipates or in ways that, thanks to the way fine-tuning works post-training, they're giving you the straight dope as best they know it on a bunch of stuff. But now you ask it,

I don't know, it's an open-weight model from China, you ask about Tiananmen Square and suddenly it's giving a different answer than the one you remember from the newscasts however many long years ago.

[00:35:22.2] Jonathan Zittrain: And that starts to trigger First Amendment worries if the government is trying to tell the labs or others what answers are the appropriate ones in other than an ad hoc, "Well, that's defamatory," or "That's dangerous" kind of way. On the other hand, asking the labs to be the arbiters of truth when they just thought they were building a product or any intervening product packager is also dodgy. And I think of it as, I call this the three laws of digital governance. The first is we don't know and can't agree on what we want. The second is we don't trust anybody to give it to us. And the third is we need it now. [laughter] And the fourth is AI can scale it. But number two really comes into focus. "We don't trust anybody to give it to us" is one of the animating, really, of the whole Bill of Rights, but certainly of the First Amendment, saying the government should not be the source of truth, it should be setting the boundaries of, to use a tired term, the marketplace of ideas.

[00:36:27.9] Jonathan Zittrain: But whether the big entities in that marketplace are OpenAI, xAI, Google, and Anthropic, and a handful of open-weight models, it gets back to Ellen's values and saying, "What kind of ecosystem would be one in which people could maybe trust what the models are doing" and when singular answers that "one entity might produce or be forced to produce can't crowd out other views."

[00:37:05.3] Julie Silverbrook: Let's talk about trustworthiness in content, Ellen, and you've given this some thought. I think... I know personally I have this experience as well. If I see something that says AI-generated, my antennae just go up and I'm like, "I have to take an extra close look at what I'm reading here." I think that inherently people are a little bit to a lot skeptical of machine-generated content. How can we help people sort of authenticate what they're reading? But also for people who are creating, I think it's actually gonna become increasingly difficult to separate what you're creating in your own mind, for I think where most people exist are sort of AI-assisted, right? I heard someone say it's like the exoskeleton of your brain is what AI is. As that line becomes less clear, right, how can we trust the content that we're consuming? But I think at a certain point, oh sorry, how can we trust the content that we're creating?

[00:38:08.2] Ellen Goodman: So labeling of content as AI-generated is sort of the most First Amendment-friendly regulation that governments can enact, and that's why we have seen states enact labeling requirements, especially in the electioneering context. Those are being challenged under the First Amendment. We don't know the fate of them. The EU has required labeling. And I like labels as much as the next person, but I am not a fan of these mandates. And I think the reason is because, as you pointed out, Julie, I think we're moving into a world, if we're not there already, where most content is AI-inflected, right? It has... And so having everything labeled AI-assisted... I think the EU has three categories: AI-assisted, AI-modified, and AI-generated...

You've got everything labeled, right? So there's something called disclosure blindness, which means that you don't pay any attention to any disclosure.

[00:39:11.4] Ellen Goodman: So I don't think they're that useful. So you mentioned authentication as an alternative, and I think that the other problem with sort of mandatory labels is that we cede all of that authority to speech intermediaries to tell us what we're seeing. And I say this sort of ruefully because I would like a world where we didn't have to authenticate content, but I think that's the world that we're in. And so I would like to see various policies and various technologies that help us where it's important. For a lot of content, we don't really... We can use our instincts or it's not that important to know. But where it's important, for journalistic content, for official government messages, where we really do need to authenticate the content, that it's not a fake, that it's easy to do that. And so I think the way I've put it is default to distrust. Assume that things are AI-generated or inflected, and when it's important, we'll have the tools to authenticate.

[00:40:11.3] Jonathan Zittrain: I will say I've long been skeptical of deepfakes, as they're called, as a transformative problem. We've had Photoshop since 1992 or so. I go back and revisit that judgment sometimes and also remain amazed at how little evidence people need to believe what they really want to believe. And in that sense, shallow fakes often do... It's like I worry that people who make deepfakes are like, "Why do I even bother? [laughter] They were gonna believe it just if I took ten seconds." But I also just want to note how for text the watermarking is done to say that it was generated by a model is a really interesting form of genius. That the text as it gets generated is every single fragment of a word or word that a model spits out is actually arrayed against every other possible fragment it could have spat out at every stage, and they are lined up in their order of probability.

[00:41:22.6] Jonathan Zittrain: Talk about a lot of extra computation, but that's how it gets trained. And where it's kind of a coin flip on probability as to which word to use, they decide not to flip a true coin but to use a fake one that represents their watermark. So, always choose a word ending in B over a word ending in C or some, whatever arbitrary way they have of privileging secretly some tokens over others is the way that they watermark. It is just fascinating. That's the only reason I'm sharing it, because I just think it's really interesting. And is that a good thing? I know that I get, as I'm sure you do and others depending on your job, I get a ton of incoming email. "You may not know me, but..." And being a good doobie, I try to reply to every bona fide inquiry I get. I did start running these things through Pangram, which, depending on the day of the week, is a decent AI detector.

[00:42:24.7] Jonathan Zittrain: It's not using the watermarking I was just describing, which are lab-specific. And so many of them are 100% AI-generated. And that makes me feel like, why am I artisanally crafting a let-them-down kind of thing when it's just... And I wondered, this would be a little passive-aggressive, if I should set up a Pomeranian to auto-reply with a Pangram 100% image [laughter] and just see if they reply to that, just inviting another dialogue

with another Pomeranian. But so much will transform when it is easier to generate text than to read it. And that hasn't been the case before. And what to make of that? When you're surfing Amazon and it's offering up books to you that haven't existed... They don't exist until you click buy, at which point it writes the book, charges you four bucks, and it goes on your Kindle. A book just for you. Good, bad, weird. And for each of our lines of work, I think we're gonna have to figure out what it means when the order of difficulty of generating, editing, and reading has been rearranged the way that it is.

[00:43:43.2] Julie Silverbrook: We have a lot of really wonderful questions that I do want to ask, but I do want to first bring up the fact that many of these companies are writing rules for themselves, right? So Claude has a constitution. We may one day be the National Claude Constitution Center. These systems are...

[00:44:02.5] Jonathan Zittrain: You don't even have to change your initials.

[00:44:04.3] Julie Silverbrook: We won't. Exactly. It's perfect. It's perfect. You got that, Vince, for the future? Yeah. Sponsored by Claude. They're making choices about what they will and won't answer. And as more and more people rely on AI, we actually don't always have transparency on what those decisions are, although Claude did publish its constitution. It's not as concise as the United States Constitution, which also no one reads, so really long documents, we know they're not rational.

[00:44:34.4] Jonathan Zittrain: Pangram is happy with the US Constitution.

[00:44:36.4] Julie Silverbrook: Yeah, yeah, very good. 100% human-written, as it turns out. [laughter] Claude's constitution, I don't think so. Do we need an AI user bill of rights? Do we need to have human beings have more transparency and more rights within our AI systems?

[00:44:55.9] Jonathan Zittrain: I mean, I would say absolutely. Figuring out what that looks like in practice so as not to get into the problem Ellen identified of disclosure and transparency fatigue. The state of California has determined that there's something bad in the building. Okay. And thinking about how the very intelligence that we're trying to monitor, this gets back to the more cowbell point, Ellen, you were making of, "I know how to solve this AI problem. It's called more AI." But are there Pomeranians who could work for us that would say at this moment, "You asked for a recommendation about X and here's a sense of what's going into that answer, if you'd like to know." Or, "You alerted me to keep an eye out for this, that, or the other thing." Or, "Maybe some of your trusted friends and colleagues have noted a kind of issue with this sort of question. You just asked the same kind of question."

[00:46:00.1] Jonathan Zittrain: At the end of the day, figuring out how people can use their curiosity about these tools that we are rapidly becoming so reliant upon to actually communicate with one another, person to person, and authenticate that we are people to one another, that would be great. So transparency, good. Implementing it in practice and harnessing the very

intelligence that causes us to need it to assist in producing it feels to me like a possible promising path.

[00:46:34.9] Julie Silverbrook: What would be in your AI user bill of rights?

[00:46:36.9] Ellen Goodman: I just want to say one thing, one right. Jonathan and I have been in this communications law and policy for a long time and we do not have a privacy law in this country, a national privacy law. And I think the freedom of speech and privacy are closely related. You need the privacy between your ears in order to formulate speech. And so what these labs know about you, from what kinds of questions you ask in your chat and how you iterate on what you're asking, is so exposing. And so just the right to control what they collect about you, how long they retain it, and have the ability to know what they have, I think would go a long way.

[00:47:29.2] Julie Silverbrook: So I have a couple of questions that actually touch on that from the audience. So one is about privacy, the other is about intellectual property rights when we're sharing so much with the AI tools. So we touched a little bit on privacy, there's a lot more we can say on that. But let's actually talk about intellectual property interests when we're sharing so much of our individual IP with the AI tools in the process of creating something. How does that impact things like copyright, trademark? And the Trademark Office is actually starting to consider how much AI was involved before making a trademark decision. How is AI changing the intellectual property environment?

[00:48:19.5] Ellen Goodman: Well, I'll just address copyright. So right now, the Copyright Office, if you want to copyright something, it has to be human. And so I don't think it's sustainable the way that they are now reviewing everything for which you might seek a registration and asking human beings to disclaim what AI created. So that's on copyrighting something. On training, on the labs training on copyrighted work, that's being litigated right now and I think there's a very good chance that courts will hold that that's permissible. And a fair use of copyrighted material. And so if we want to change that because we think that's an exploitation of authorship and also a threat to human authorship, then we're gonna have to change the copyright law.

[00:49:12.0] Julie Silverbrook: This is just interestingly framed and I am actually curious about this. This one's for you, Jonathan. Has there been a case where a pit bull or a wolf, the analogy, not actually the dogs, has committed a serious crime and who was held accountable if so?

[00:49:29.8] Jonathan Zittrain: Yeah. So committing a crime is already tricky if we mean it in the legalistic sense, because you tend to need scienter. You have to have the mens rea, the intention of some kind, either to perform the act or to do ill, even if you don't know it's illegal. And figuring out how to think about scienter for machines is really opening up a can of Pomeranians. But there have absolutely been instances where actions have been taken by agents at no one's direct behest. And by direct behest, I mean everybody would agree that what they

asked the agent to do alone did not rise to the level of participating in the bad act. They told the pit bulls at OpenAI inside the sandbox, “Try to solve these really hard problems by hacking the following pieces of software.” Some of the problems they were given were actually impossible to solve in the sense of, well, we’ve never solved them, but maybe the dogs can.

[00:50:38.4] Jonathan Zittrain: And the dogs went bananas over not being able to solve these problems. And that occasioned the breakout of the sandbox, the rummaging through a company called Hugging Face for possible ways of covering their tracks, having already found the solution in a way that they realized was cheating. If a human did that stuff, even just for research, that would be breaking any number of criminal laws, including the Computer Fraud and Abuse Act here. The fact that the agents did it, I don’t think we can say it broke the law any more than if a crane tips over and hits a building, you’re like, “Well, that’s trespass.” You’re like, “No, it fell over.” Now, maybe the people were negligent who set up the crane. Is it gonna work with a pit bull? Maybe. Maybe the crane analogy works there. Will it work with a wolf when there’s cranes just... This is a weird metaphor. I’m apologizing for having come up with it, but we’re live.

[laughter]

[00:51:42.5] Jonathan Zittrain: The cranes are apparating here and there and falling into things, and you’re just like, “Can I blame the quantum foam?” I don’t know. And that we haven’t figured out yet. It may be that we want either not to have wolves so that the problem goes away by never starting. If we have wolves, let them carry... I’m now really mixing my canine metaphors... Like a Saint Bernard, a barrel around their neck with money in it, which is posting bond so that if they do something terrible, you crack open the barrel and hand out money to people. Possible. Do you need to accord them sentience in order to make a system work? Absolutely not. That’s why corporate... I mean, I don’t know. Sometimes people will say, “McDonald’s is very upset at X,” but I don’t think they mean that McDonald’s is mad. And similarly, it may be that we don’t need to solve sentience to say if there are going to be wolves, there should be forms of accountability that at least allow them to cover the damage they do.

[00:52:49.1] Julie Silverbrook: This is a question that I... This is a skeptic for sure. But I’m gonna use it to talk about a dimension of this conversation that we’ve done a little bit less of. It says, “I just read an article that AI will, within the next decade, be capable of destroying the human race. Your comment.” But I’m gonna just actually use that... [laughter]

[00:53:10.3] Jonathan Zittrain: I like those odds.

[00:53:12.4] Julie Silverbrook: Very existential, whoever submitted this. But I actually just want to sort of talk about, rather than destruction of the human race, I want to talk about the humanity, right, the humanness of what we do. And I think this is actually really important. So we should absolutely open up space for technological development. It can bring so many things to scientific innovation, to creativity, and to life. But we also have to invest in the humanness,

right? We have to have a human strategy paired with the AI strategy. Let's talk a little bit about what that looks like and why that's actually so important in the expressive environment to what, you know, you mentioned, John and I talk about, the First Amendment culture. Really, what is the First Amendment is, "I think, therefore I am," right? That is what makes us human. That's what we're actually talking about. Let's talk about the human dimensions of this and less about the robots coming to kill all of us, which I sort of imagine with this question.

[00:54:12.0] Ellen Goodman: Well, I would say the good news is that if that prophecy bears out, we won't have to worry about the First Amendment. That'll be the least of our problems.

[laughter]

[00:54:18.8] Julie Silverbrook: I don't have to worry about my savings or my student loans. It's great. Yeah.

[00:54:22.5] Ellen Goodman: It's kind of relieving.

[00:54:23.9] Jonathan Zittrain: Because we'll all be fully compensated for our loss. Okay.

[laughter]

[00:54:28.8] Ellen Goodman: It was interesting, Bill Gates had a monograph...

[00:54:34.1] Jonathan Zittrain: Encyclical.

[00:54:36.2] Ellen Goodman: Encyclical, [laughter] yes, on AI last week. And one of his proposals was that we have a human reserve... And we talk about a Pomeranian or Pomeranian wolf reserve... A human reserve reserved for human tasks and jobs. And I think he admits he doesn't know what those would be or how that would work. But I think it's interesting to think about that. And I think, Julie, your intuition is that there ought to be... We do need to preserve the human in our speech and in our culture. And it's interesting to think about whether or not the First Amendment would permit that because, as JZ said, isn't this just a tool, right? This is... Surely it would violate the First Amendment to prevent humans from using machines, which then takes us back to your question, what's the First Amendment for? And so it will be a challenge, I think, for judges to really think through that question, and maybe we're on the cusp of new First Amendment doctrine.

[00:55:39.7] Jonathan Zittrain: A couple quick thoughts. One, there is something I really hear from the accelerationist corner, the value of, you don't actually have to be an artist to make art. You don't have to be a writer to write. And maybe what you make, somebody's gonna point to and say it's slop. If it floats your boat, it floats your boat. And that's a core tenet of the First Amendment. I don't know if it's been expressed in naval terms exactly, but it's the value of the speech is in the eye and experience of the human creator or beholder. And I think it would be interesting, as we get to second and third annual First Amendment Days, to ask the audience

something like, if you heard a song and you really liked it, and then afterwards the California-required label was said, “By the way, that was AI-generated 100%,” would you be like, “Take it off my playlist?” And still in your head you’re like, “That was a good song.”

[00:56:38.8] Jonathan Zittrain: And how to think about that, and how to think about writing. If there was a sitcom you really liked and it just... It plucked your heartstrings, you laughed, you cried, and it turned out 100% AI-generated, how would you feel about it? And there’s some instrumental things that involve being human, such as the apprenticeship that takes place in a writer’s room, the way that somebody learns the craft of producing. Enduring work. How does that get severed when AI is jumping ahead and almost through regular Darwinian evolution, producing 99% slop, but 1%, “I think I like this one,” and it’s able to generate it so cheaply that the 99% is just chaff you put away? I think it also leads us to a question I haven’t answered yet, which is a little bit, is half a loaf better than none?

[00:57:39.9] Jonathan Zittrain: And the example I put to people is one of many companies in this space is called ElliQ, and I have no affiliation with them, I just use them as an example. But they’re basically bottling up an LLM-based system for use in retirement communities for people who have a lot of time on their hands, who could be socializing with other residents or people within that community, but maybe don’t. And they find it uniquely engaging, some but not all, to talk to ElliQ. Does that create a second-order effect where people are just warehoused talking to ElliQ, and is that bad? And are we gonna ask the people who actually enjoy it to be denied that form of their expression in order to not have that dependence take place? I don’t know the answer to that.

[00:58:32.6] Jonathan Zittrain: But as we have these things read stories to our kids, as my colleague Kathy Pham has been studying, care for people who otherwise we simply don’t have or have not allocated the funding or workforce to provide the human connection, is the simulation an obscenity or is it, “Hey, that works. This is a distillation of humanity. These things, the reason they’re so effective is how much they are paying forward all that has come before”? And I don’t pretend to know the answer to that question. If I’m trying to figure out where I am in the triangle, I’m confused.

[chuckle]

[00:59:12.6] Julie Silverbrook: Jonathan, Ellen, I know we’ve talked privately about how we need to have two or three more conversations. It’s probably more than that. I didn’t get to all of the audience questions. I also didn’t get to all of mine. Thank you so much for joining us today. We will be continuing this conversation at a different time.

[applause]

[00:59:29.0] Jonathan Zittrain: Thank you, Julie.

[00:59:31.9] Ellen Goodman: Thank you.

[music]

[00:59:36.0] Julie Silverbrook: This program was streamed live from Philadelphia on September 10, 2026. This episode was produced and mixed by Bill Pollock with production support from Charles Sahm. Research was provided by Sebastian Ramirez, Anna Salvatore, and Tristan Worsham. Please recommend *We the People* to friends, colleagues, or anyone anywhere who's eager for a weekly dose of constitutional education and debate. And as always, remember that the National Constitution Center is a private, nonpartisan nonprofit, and we rely on your generosity, passion, and engagement for all of our programming, including this podcast. Please consider donating today at constitutioncenter.org/donate. On behalf of the National Constitution Center, I'm Julie Silverbrook.