UNITED STATES OF AMERICA
Federal Trade Commission
WASHINGTON, D.C. 20580

Office of Commissioner
Alvaro M. Bedoya

"Early Thoughts on Generative AI"

Prepared Remarks of Commissioner Alvaro M. Bedoya, Federal Trade Commission
Before the International Association of Privacy Professionals
April 5, 2023

Thank you for that kind introduction. I want to thank the IAPP for having me, and I want to thank my team, specifically Aaron Rieke and Danielle Estrada, for being my thought partners in this work. Before I begin, I'll say that I speak for myself, not the Commission, its staff, or any of my fellow commissioners.

There have been a lot of headlines about generative AI these last few months. "This is the end of modern English education." "This will result in mass unemployment." "It is sentient." "It poses an existential threat to humanity."[1]

I want to slow things down and take a sober look at what's been happening in AI. And while I won't cover some critical subjects, I want to offer you three reflections:

- First, you do not need to believe the most *breathless* predictions about this technology to appreciate that it *can* induce awe and *wonder*.

- Second, you do not need to believe the most *dire warnings* about this technology to understand that it presents new and *substantial* risks for the American public.

- Third, I want to offer my own warnings about how those risks map onto our American system of consumer protection.

I want to share this with *you* because I know that many of you aren't *just* privacy professionals, but also de facto ethicists and in-house AI experts.

---

[1] *See e.g.* Daniel Herman, *The End of High-School English*, ATLANTIC (Dec. 9, 2022), https://www.theatlantic.com/technology/archive/2022/12/openai-chatgpt-writing-high-school-english-essay/672412/ (for the claim AI will be the end of modern English education); Steven Greenhouse, *US Experts Warn AI Likely to Kill Off Jobs – and Widen Wealth Inequality*, GUARDIAN (Feb. 8, 2023), https://www.theguardian.com/technology/2023/feb/08/ai-chatgpt-jobs-economy-inequality (for discussion of claims that AI will result in mass unemployment); Sarah Palmer & Sophia Khatsenkova, *'I want to be alive': Has Microsoft's AI chatbot become sentient?*, EURONEWS.NET (Feb 18, 2023), https://www.euronews.com/next/2023/02/18/threats-misinformation-and-gaslighting-the-unhinged-messages-bing-is-sending-its-users-rig (for discussion of claims that AI has become sentient); Thomas Barrabi, *Elon Musk Warns AI 'One Of Biggest Risks' to Civilization During Chatgpt's Rise*, N.Y. POST (Feb. 15, 2023), https://nypost.com/2023/02/15/elon-musk-warns-ai-one-of-biggest-risks-to-civilization/ (for the claim that AI is an existential threat to humanity).

1. Zeus in a Three-Piece Suit

DALL-E renderings of "French bulldog climbing a skyscraper in the clouds" and "a tiger wearing a backpack shooting red lasers"

I suspect most of you have a story about how you discovered generative AI. Here's mine.

A few months ago, I created an account on DALL-E. I thought I'd see if I could use it to entertain our two toddlers. We have a little French bulldog at home. So, you know, I put her on some skycrapers. My son likes a little more drama, so I made him a tiger wearing a backpack shooting red lasers. I thought it was neat, but I wasn't too impressed. I filed them away for a slow day at home. Didn't think much of it.

Then I saw a note on social media saying that the image generators do interesting stuff when you ask them to render images of Greek gods. Then I saw something else that said: Ask the generators to render the images in the style of a famous artist. So the next time I was in front of a computer, on a whim, I asked it to render an image of "Zeus in a Three-Piece Suit" by Edward Hopper, the famous American oil painter. And I got… this.

DALL-E rendering of "Zeus in a three-piece suit by Edward Hopper"

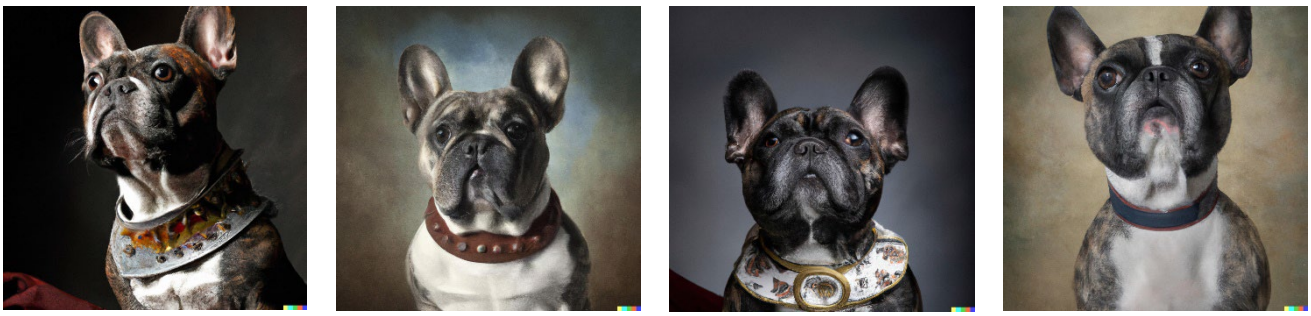
Hopper had an uncanny ability to communicate existential dread with ostensibly neutral images. And he had a real skill with dim, interior lighting. Yet despite that skill, Hopper was never particularly good at faces. Looking at it, I thought: This makes me feel the way a Hopper does. It knocked me out.

Edward Hopper, "Approaching a City" (1946) & "New York Movie" (1939)

So I thought: Okay, let's see what we can do with this! I always liked portraits by the Old Masters, but I thought it was funny that the sitters always wore these big frilly white lace collars.

Rembrandt van Rijn, Portrait of a Gentleman (c. 1634) & Portrait of a Woman (1635)

So I thought, let's have the Old Masters paint our bulldog. I asked DALL-E to render "A brindle French bulldog wearing a white lace collar painted by Rembrandt." And I got these guys.



DALL-E rendering of "A brindle French bulldog wearing a white lace collar painted by Rembrandt"

This was disappointing. Where's the frilly collar? They also don't look anything like Rembrandts. Maybe Rembrandt just doesn't work with the bulldogs. So I swapped in Hans Holbein the Younger, Henry VIII's court portraitist who painted these head on, very realistic portraits, usually against a monotone background – easier, right?

Hans Holbein the Younger, Portrait of Henry VIII (1540) &
Detail from Portrait of Christina of Denmark (1538)

Not really. It still looked like someone had taken a photo of a bulldog dressed up for a renaissance painting.


DALL-E rendering of "a brindle French bulldog wearing a while lace collar
painted by Holbein the Younger"

So I thought: Maybe the issue is that it's a bulldog. Maybe there are too many images of a bulldog in one pose, and maybe – given the recent popularity of the breed – the vast majority are high-definition photos rather than drawings or paintings. Maybe that unusually uniform training data would skew DALL-E to render the images in a particularly uniform way.

I needed an animal that doesn't have a "good" side – and that's never been a social media darling. So I chose the opossum, a marsupial native to the Americas.

An opossum on top of a fence. Photo by Sergey Yarmolyuk (CC-BY-SA 4.0)

So I rendered the opossum in that formal white collar. First, in Rembrandt. And I got this.

DALL-E rendering of "a possum wearing a white lace collar painted by Rembrandt"

And *that*, that looks like Rembrandt; the dark background, the sfumato, the soft transitions between colors.

I tried Seurat. I tried Van Gogh – and actually, this looks nothing like a Van Gogh. But we'll come back to that. Instead I decided to go modern. I tried the photographer Annie Liebowitz. Then the modern portraitist Kehinde Wiley.

DALL-E renderings of "a possum wearing a white lace collar painted by Georges Seurat" (top left); "...painted by Van Gogh" (top right); "...photographed by Annie Liebowitz" (bottom left); "...painted by Kehinde Wiley" (bottom right)

I couldn't stop thinking about these images. Eventually, a few thoughts brought me back down to earth. The first is that I selected the images I thought were most compelling in response to my prompts. I didn't show you the others. Some of these other ones are terrible!

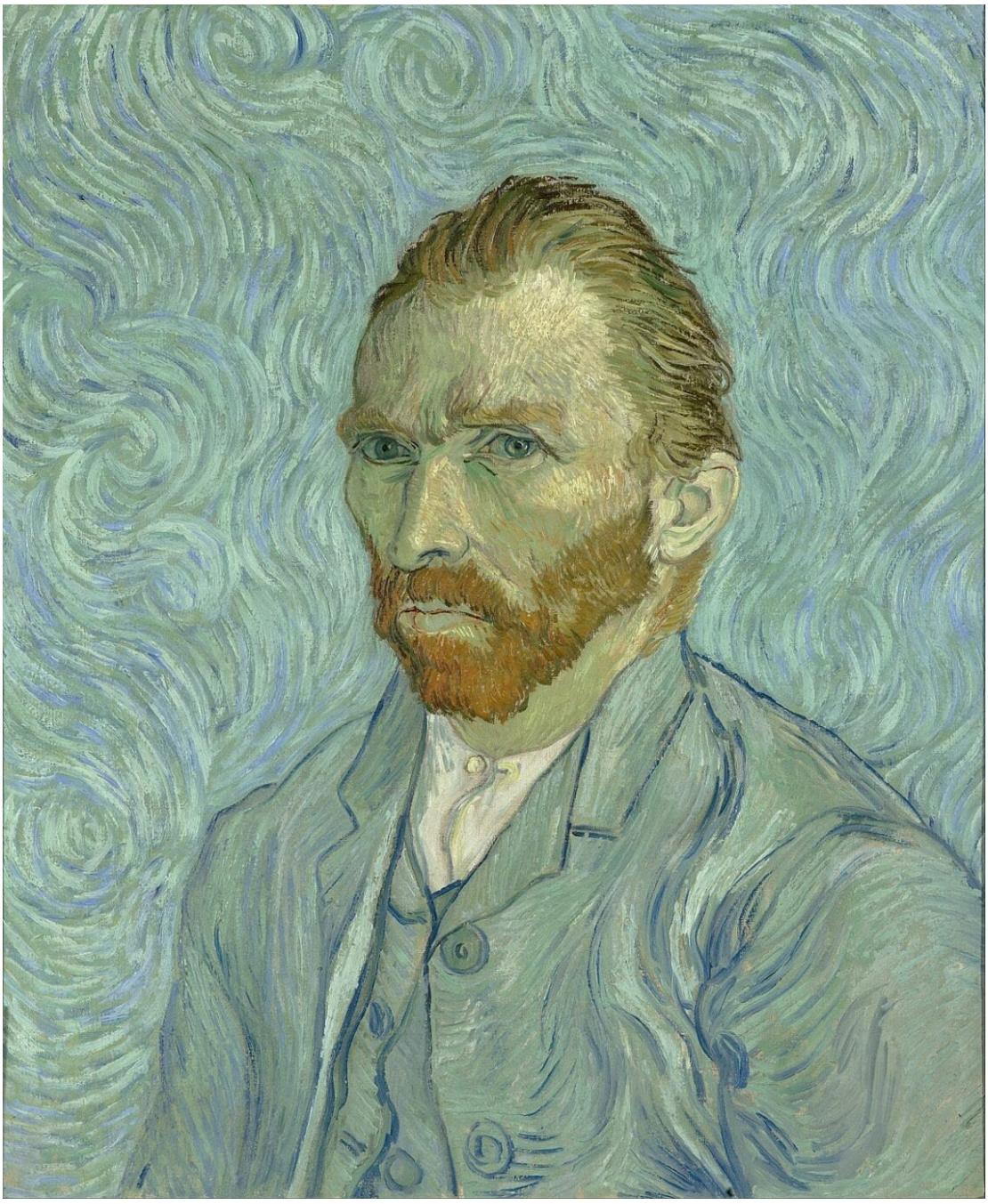
This is the good Hopper… some of these other ones are soulless, this one is way too bright, and what the heck is this? Zeus wearing a two-tone tie and… suit shorts?

DALL-E renderings of "Zeus in a three-piece suit by Edward Hopper"

Here's the good Seurat, and again -- these are passable, but this is just terrible.

DALL-E rendering of "a possum wearing a white lace collar painted by Georges Seurat"

And then what about Van Gogh: Why on Earth can it not do Van Gogh? Van Gogh has rich backgrounds, defined lines in sharply contrasting colors.



Vincent van Gogh, Self-Portrait (1889)

These have nothing to do with a Van Gogh. And this -- this is a rat. And why does DALL-E keep giving the rats little bow ties?

DALL-E renderings of "a possum wearing a white lace collar painted by Van Gogh"

And let's be honest here. If you're an art student, and you turn *this* [*bottom right, above*] in for your Van Gogh study… you've got some hard choices ahead of you. That's the first thing: I made the model out to be more than it was by focusing on its optimal performance, not its standard performance.

Second, I realized while DALL-E's renderings were impressive, the true *creativity*, the *genius*, *that* was in the prompts and the underlying training data. DALL-E *mimicked that genius*.

The genius is in how Rembrandt mastered the ability to capture something more accurately by painting it *less precisely*. DALL-E mimicked that. The genius is in how Kehinde Wiley relentlessly studied the Old Masters, perfected their technique, and experimented with transposing that technique into modern subjects. DALL-E mimicked that.

Rembrandt van Rijn, Self-Portrait (1659) & Detail

Jean-Auguste-Phillipe Ingres, Napoleon on His Imperial Throne (1806) & Kehinde Wiley, Ice T (2005)

I'm not trying to downplay what the creators of generative AI have accomplished. I *am* trying to say that it doesn't work from a clean slate; it doesn't produce content out of thin air. *And*, *at the same time, it is full of wonder*.

Too often, people in my role fail to recognize the beauty and wonder in the technologies we regulate. Instead, we focus on bureaucratic line drawing. I don't want to fall for that.

2. Surprising, Inexplicable, and "A Little Bit Scary"

That said, I do think that generative AI presents fundamentally new dynamics for a consumer technology.

What are those dynamics?

- First, developers of generative AI are *surprised* by their models' performance, and often cannot predict it.

- Second, these developers often cannot *explain* significant aspects of that performance.

- And third, many of these developers insist that they find the technology "a little bit scar[y]."[2]

Let's start with *surprise*. And let me add that, much as I love opossums, I'll focus on large language models for the remainder of this conversation.

At the most basic level, LLMs are a collection of simple operations, taking numerical inputs and combining them with certain weights. Yet from this truly simple foundation come astounding "emergent" behaviors, in other words, behaviors which cannot be easily reduced to or explained by their constituent parts.

One example. GPT-3 and GPT-4 are known for their translation skills. Turns out that this skill came as a surprise in earlier versions of the model.

Before training GPT-2, the engineers working on it had purposefully removed non-English websites from its training set. Yet, when they tested GPT-2 on some standard English-French translation benchmarks, they realized that it did reasonably well on a few tasks, even beating a number of unsupervised machine translation models. The engineers eventually figured out that the model had been trained on just 10 megabytes of French data that had snuck past the filter – that is 500 times less than the amount of French language that had previously been used to train unsupervised machine models.

[2] ABC 7 Chicago, *CEO Behind Chat GPT-4 Says He's 'a Little Bit Scared' by AI*, YOUTUBE, at 0:52 (Mar. 17, 2023), https://www.youtube.com/watch?v=_PBh9BzMpGM.

And yet, GPT-2 could beat some of those models.[3]

That was a few years ago. The latest papers insist that we cannot effectively predict what new abilities will appear, the scale at which those abilities will appear, or even what's ultimately possible.[4]

But this technology is not just surprising — it's also uniquely *inexplicable*.

I just mentioned translation. It turns out that modern large language models can also play chess. That sounds simple. What's fascinating is that there does not appear to be a consensus among technical experts about *how the model does this*.[5]

Is it relying on trends and statistics? Is the model figuring out that after hearing about moves X, Y, and Z, a player typically says something about a queen taking the rook? Or is it somehow tracking the state of play using a "world model," despite just being a word prediction engine that was never trained to play chess?

The top minds in our field don't yet agree upon an answer. The top-line is that today, we can't just "pop open the hood" to see how these models work. Today, they are among the darkest of the black boxes.

Last, let's turn to the idea that the creators of this technology are "a little bit scared" to quote the CEO of ChatGPT.[6] Personally, and I say this with respect – I do not see the existential threats to our society that others do. Yet when you combine these statements with the unpredictability and inexplicability of these modfels, the sum total is something that we as consumer protection authorities have never reckoned with.

Let me put it this way. When the iPhone was first released, it was many things: a phone, a camera, a web browser, an email client, a calendar, and more.

Imagine launching the iPhone – having 100 million people using it – but not knowing what it can do or *why* it can do those things, all while claiming to be frightened of it.

That is what we're facing today. So we need to think quickly about how these new dynamics map onto consumer protection law.

---

[3] Alec Radford, et al., *Language Models are Unsupervised Multitask Learners*, 1(8) OPENAI BLOG (2019), https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[4] Samuel R. Bowman, Eight Things to Know about Large Language Models 2-4 (unpublished manuscript), https://cims.nyu.edu/~sbowman/eightthings.pdf ("[Section] 2. Specific important behaviors in LLM tend to emerge unpredictably as a byproduct of increasing investment").

[5] Kenneth Li, *Do Large Language Models Learn World Models or Just Surface Statistics?*, GRADIENT (Jan. 21, 2023) https://thegradient.pub/othello/.

[6] ABC 7 Chicago *supra* note 2 ("I think people should be happy that we're a little bit scared of this.")

3. The Adult (Human) in the Room

And so, I'll offer four observations that double as notes of caution.

- First, generative AI is regulated.

- Second, much of that law is focused on impacts to regular people. Not experts, regular people.

- Third, some of that law demands explanations. "Unpredictability" is rarely a defense.

- And fourth, looking ahead, regulators and society at large will need companies to do much more to be transparent and accountable.

Let's start with that first point. There is a powerful myth out there that "AI is unregulated." You see it pop up in *New York Times* op-ed columns,[7] in civil society advocacy, and in scholarship. It has a powerful intuitive appeal — it just *sounds* right. How could these mysterious new technologies be regulated under our dusty old laws?

If you've heard this, or even said it, please take a step back and ask: Who does this idea help? It doesn't help consumers, who feel increasingly helpless and lost. It doesn't help most companies. It certainly doesn't help privacy professionals like you, who now have to deal with investors and staff who think they're operating in a law-free zone.

I think that this idea that "AI is unregulated" helps that small subset of companies who are uninterested in compliance. And we've heard similar lines before. "We're not a taxi company, we're a *tech* company." "We're not a hotel company, we're a *tech* company." These statements were usually followed by claims that state or local regulations could not apply to said companies.

The reality is, AI *is regulated*. Just a few examples:

- Unfair and deceptive trade practices laws apply to AI. At the FTC our core section 5 jurisdiction extends to companies making, selling, or using AI.[8] If a company makes a deceptive claim using (or about) AI, that company can be held accountable. If a company injures consumers in a way that satisfies our test for unfairness[9] when using or releasing AI, that company can be held accountable.

---

[7] Rep. Ted Lieu, *I'm a Congressman Who Codes. A.I. Freaks Me Out*, N.Y. TIMES (Jen. 23, 2023), https://www.nytimes.com/2023/01/23/opinion/ted-lieu-ai-chatgpt-congress.html.
[8] Michael Atleson, *Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale*, FTC: BUS. BLOG (Mar. 20, 2023), https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale.
[9] *Policy Statement on Unfairness*, FTC (Dec. 17, 1980), https://www.ftc.gov/legal-library/browse/ftc-policy-statement-unfairness.

- Civil rights laws apply to AI. If you're a creditor, look to the Equal Credit Opportunity Act. If you're an employer, look to Title VII of the Civil Rights Act. If you're a housing provider, look to the Fair Housing Act.

- Tort and product liability laws apply to AI. There is no AI carve-out to product liability statutes, nor is there an AI carve-out to common law causes of action.

AI *is* regulated. Do I support stronger statutory protections? Absolutely. But AI does not, today, exist in a law-free environment.

Here's the second thing. There's a back-and-forth that's playing out in the popular press. There will be a wave of breathless coverage – and then there will be a very dry response from technical experts, stressing that no, these machines are not sentient, they're just mimicking stories and patterns they've been trained on. No, they are not emoting, they are just echoing the vast quantities of *human* speech that they have analyzed.

I worry that this debate may obscure the point. Because the law doesn't turn on how a trained expert reacts to a technology – it turns on how regular people understand it.

At the FTC, for example, when we evaluate whether a statement is deceptive, we ask what a reasonable person would think of it.[10] When analyzing unfairness, we ask whether a reasonable person could avoid the harms in question.[11] In tort law, we have the "eggshell" plaintiff doctrine: If your victim is particularly susceptible to an injury you caused, that is on you.[12]

The American Academy of Pediatrics has declared a national emergency in child and adolescent mental health. The Surgeon General says that we are going through an epidemic of loneliness.[13]

I urge companies to think twice before they deploy a product that is designed in a way that may lead people to feel they have a trusted relationship with it or think that it is a real person. I urge companies to think hard about how their technology will affect people's mental health – particularly kids and teenagers.

Third, I want to note that the law sometimes demands explanation – and that the inexplicability or unpredictability of a product is rarely a legally cognizable defense.

---

[10] *Policy Statement on Deception*, FTC at 2 (Oct. 14, 1983), https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf ("The Commission believes that to be deceptive the representation, omission or practice must be likely to mislead reasonable consumers under the circumstances.")

[11] *Policy Statement on Unfairness supra* note 9 ("To justify a finding of unfairness the injury…it must be an injury that consumers themselves could not reasonably have avoided.")

[12] *See* 2 JACOB A. STEIN, STEIN ON PERSONAL INJURY DAMAGES TREATISE § 11:1 (3d ed.)

[13] Jena McGregor, *This Former Surgeon General Says There's a 'Loneliness Epidemic' and Work is Partly to Blame*, WASH. POST (Oct. 4, 2017), https://www.washingtonpost.com/news/on-leadership/wp/2017/10/04/this-former-surgeon-general-says-theres-a-loneliness-epidemic-and-work-is-partly-to-blame/.

What do I mean by that?

Looking solely on laws that the FTC enforces, both the Fair Credit Reporting Act and the Equal Credit Opportunity Act require explanations for certain kinds of adverse decisions.

Under our section 5 authority, we have frequently brought actions against companies for the failure to take reasonable measures to prevent reasonably foreseeable risks. And the Commission has historically not responded well to the idea that a company is not responsible for their product because that product is a "black box" that was unintelligible or difficult to test.

I urge companies who are creating or using AI products for important eligibility decisions to closely consider that the ability to explain your product and predict the risks that it will generate may be critical to your ability to comply with the law.

Fourth and last, I want to end on a call for a maximum of transparency and accountability.

I recently saw that the technical report accompanying the GPT-4 rejected the need to be transparent about the building blocks of the technology. It says:

"Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[14]

This is a mistake. External researchers, civil society, and government need to be involved in analyzing and stress testing these models; it is difficult to see how that can be done with this kind of opacity.

4. Focusing on Threats *Today*

I keep thinking about the now-infamous survey conducted by Oxford University last year that found that the median expert gave a 5% chance that the long-run effect of advanced AI on humanity would be, and I quote, "extremely bad (e.g. human extinction)."[15] I also keep thinking about the GPT-4 white paper's focus on "*safety*."[16]

I'm worried that inchoate ideas of existential threats will make us – at least in the short and medium term – much *less* "safe." I'm worried that these ideas are being used as a reason to provide less and less transparency. And I worry that they might distract us from all the ways that AI is *already* being used in our society today.

---

[14] OPENAI, GPT-4 TECHNICAL REPORT (2023) at 2, https://cdn.openai.com/papers/gpt-4.pdf.
[15] Katja Grace et al., *Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts,* 62 J. A.I. RSCH. 729, 733 (2018), https://jair.org/index.php/jair/article/download/11222/26431/.
[16] OPENAI *supra* note 14 at 41-78 (GTP-4's system card focuses almost entirely on the safety implications of the model).

Automated systems new and old are routinely used *today* to decide who to parole, who to fire, who to hire, who deserves housing, who deserves a loan, who to treat in a hospital – and who to send home. These are the decisions that concern me the most. And I think we should focus on them.

Thank you for your time today. It's good to see you all.