*Subcommittee on Human Rights and the Law*

Hearing: "Artificial Intelligence and Human Rights"

# Written Statement of Aleksander Mądry[1]

June 13th, 2023

Chairman Ossoff, Ranking Member Blackburn and Members of the Committee, thank you for inviting me to testify. Much has already been said and written about how AI may transform society, both about the opportunities and risks—from AI's potential to enhance our productivity, creativity, and overall quality of life to its ability to perpetuate discrimination, drive economic inequality, and pose an existential risk.

I will not reprise those conversations here. Instead, I will focus my testimony on one issue that I find particularly salient, time-sensitive and extremely worrisome: *how AI could erode central tenets that enable our society to function, including our ability to carry out democratic decision-making.*

Specifically, I will discuss how AI is poised to fundamentally transform mechanisms for the dissemination and understanding of information, and the unsettling implications of those changes. I will also sketch out what could be done to mitigate these emerging risks.

## How will AI transform the information ecosystem?

Changes in information technologies—whether the invention of the printing press, the advent of e-mail, or the emergence of social media—do not just make information more accessible, they fundamentally change the dynamics of information sharing and acquisition. While we are still dealing with the transformations in this space brought to us by email and social media, there is already a new transformation afoot—a transformation fueled by recent developments in AI that is likely to be more consequential than anything we have been experiencing recently.

With the advent of AI—especially the newest wave of generative AI—anyone who can use a chatbot is in a position to become a "trusted source"—a *highly personalized* source, in fact. Indeed, as more of what we see becomes generated and disseminated by AI, the lines between humans and bots are becoming blurred. We need to start to be more wary than ever about how information reaches us, its trustworthiness and its ability to persuade us.

More precisely, AI is changing the information-delivery landscape in three key ways:

(a) It enables the creation of content—written text, photos and, soon, videos—that seems extremely realistic.

(b) The language produced by Large Language Models (LLMs) like chatGPT or Google Bard can seem natural and highly persuasive, in no small part since we are wired to believe that such speech can come only from humans.

(c) It makes the creation of such content cheap and broadly accessible—even to parties with little if any technical expertise—making it frighteningly easy to deploy it at scale.

[1] I have recently started a professional leave from MIT, which I am spending at OpenAI. I am providing this testimony solely in my personal capacity and as an MIT faculty. I am not in any way representing OpenAI.

We are already seeing early adoption of generative AI in our information sphere, from art [10], to copywriting [7], to political ads [13], but these are just a tip of the iceberg. We will see much, much more very soon. The onset of this technology brings with it a whole spectrum of risks and potential harms. I will highlight just a few of them below.

**Enhancing "traditional" cybercrime.** One immediate impact of the newest wave of generative AI is that "traditional" spam and phishing campaigns are even easier to conduct. What previously required careful photo editing and writing (as well as some non-trivial human involvement) now only requires a few clicks. The recent use of an AI-generated fake image of a fire near the Pentagon is just one illustration of that [9].

Also, the fact that generative AI can convincingly impersonate a human online poses a fundamental challenge to our existing mechanisms for protecting our information infrastructure, public discourse and governance. After all, the bot detection and moderation algorithms that our online discussion platforms use—whether they be Internet forums, newspaper comment sections, or Twitter—tend to rely on some kind of "prove that you are human" tests. How will these platforms cope with malicious parties that can field swarms of sophisticated, AI-driven bots that are able to breeze through such tests?

**"Spear-phishing" and personalized blackmail.** The enhancement of the "traditional" deception is, however, just the beginning. AI's unique ability to create content that is both convincing and personalized means that, for example, phishing will no longer need to involve generic emails sent out to thousands of recipients, hoping someone will get duped. Instead, we will have "spear-phishing," where both the message *and* the whole conversation that ensues are *fully automated* and *customized* to you.

In fact, there is a very real possibility that a new kind of blackmail scheme will emerge. In such a scheme, someone's photo from social media is edited to depict them in a compromising situation, and then they are threatened that the edited photo will be made public unless they pay up. How many of us would not pay to simply make the problem go away? Thanks to AI these kinds of schemes can now be executed (again) *fully automatically*, *cheaply* and *at scale*.

In addition—as one of the other witnesses has experienced herself [3]—the AI-fueled ability to impersonate the voice of just about any person enables a whole new array of scams [12]. As the ability to generate video with AI improves, other troubling possibilities such as targeted AI-generated explicit content [2] will become an even more acute problem too.

**Personalized persuasion at scale.** This expansion of the cybercrime toolkit is hardly the only worrisome consequence though. Indeed, AI is bound to transform how we think about any information campaign—be it ideological, political or commercial. Specifically, such campaigns will no longer need to rely solely on the promoted message to go viral. Instead, they can be fielded with generative AI and the promoted messaging might reach its intended audience *individually* and in a *highly personalized* manner. So, it will not be about some post that came across your social media timeline. Rather it will be about a Facebook "friend" that you met online. Friend who is actually an AI-driven agent impersonating a human. Friend that only subtly weaves in political commentary or product endorsements or any other messaging in between your engaging conversations about sports, movies or favorite food.

Similarly, instead of trying to corral a critical mass of people to campaign for a cause—whether on social media, via direct calling, or letter-writing—a single actor can field a campaign by themselves, using generative AI-driven bots in place of people. A campaign that is *equally effective* (thanks to the sophistication of these bots) but needs neither any buy-in from the broader population nor even comparable resources. As far as I know, as of now, this would all be legal too.

**Automated creation of addictive content.** AI doesn't just produce content that mimics reality and appears human-like and personalized—it can also make this content *personable*. There is a lot of information about our habits, preferences, hobbies and values that can be gleaned from sources such as our social media accounts. This could make interacting with AI not only attractive and persuasive but also addictive to us. After all, loneliness and an unmet need for some kind of intimacy with others are a growing problem in our society [8], and the kind of focus, "fit" and "care" such AI-driven "friends" would seem to exhibit could be extremely alluring.

This aspect of AI could (and, I hope, will) play a positive role too [1]. But imagine the power someone who is able to deploy such AI-powered agents could have over us, especially at scale. What if that power gets abused? What if these capabilities are harnessed to supercharge the "attention economy" that already drives much of our social media and online commerce? What would this mean for our productivity and long-term happiness? How do we feel about having our children being exposed to all of that?

**Eroding trust in information and written (or audio-visual) records.** Thanks to AI we are entering the era when *any* record could plausibly be faked. How does this affect our collective discourse as well as the legal and governance system? After all, we are a society whose foundations rely on the veracity and binding of such records—think contracts, deposition recording, or video evidence in criminal cases—and this reliance will only increase as more of our critical interactions occur in the digital sphere. How does our society adapt to such a tectonic shift?

## What can we do?

The concerns I have outlined above may paint a rather bleak and, potentially, daunting landscape. But there is much we can (and should) do here. Specifically, we need a combination of technical solutions and policy actions that will reinforce each other. After all, policy can help drive the development and implementation of technical remedies, and technical innovations can, in turn, unlock new policy options. Let me describe some of these below.

### Technical solutions

On the technical front, we need tools that can help humans judge the authenticity of content—to understand the extent to which it was generated by a human and/or AI. These tools can take a variety of forms (and for many of them we already have proof-of-concept prototypes):

**Watermarking and deepfake detection tools.** One promising idea for ensuring the authenticity of content is "watermarking"—that is, placing an imperceptible "signature" in generated content that makes clear AI was used. This watermark can then be detected by any content consumer. Researchers have developed prototypes of watermarking systems, both in the context of large language

models [4] and image generations models [14]. Much more work is needed, however, to make them sufficiently robust and then policies might be needed to drive their adoption too. Also, like all such technologies, there would likely be an "arms race"—tools will be developed to evade the watermark system and improved techniques will be needed to respond to that.

Watermarks need to be placed in documents directly by the AI providers, but there is also a line of work on detecting AI-generated content in the absence of cooperation from the developers of a given AI model [6]. Of course, this lack of cooperation makes it easier for malicious actors to thwart these detection techniques, causing the corresponding "arms race" to be much more challenging.

**Protection against unauthorized AI-powered content editing.** Another problem that technology can help address is unauthorized AI-powered content editing—that is, the ability to use AI-powered editing tools to manipulate content against the wishes of its creators or people depicted in it. (Think, for example, of the personalized blackmail scheme described earlier, which involved a malicious party manipulating photos the victim had published on social media.) Could we develop a way for users to protect the photos they put online, to make it impossible—or, at least, much harder—to modify using AI? It turns out that such an "immunization" capability is a possibility [11] but, again, much more work is needed.

**Provenance certification techniques.** Beyond detecting AI-generated content, tools may be needed to *prove* the authenticity of content. This could involve, for example, leveraging cryptographic tools to provide automatic certification of the authenticity or provenance of a given document by tracing it to the exact primary source that created it (e.g., the person who took a given photo). When such a technology is broadly available, content might be presumed to be fake unless verification proves it to be real.

However, just to reiterate: no matter how work on such tools proceeds, these tools will *not* be a panacea. They will be neither perfect nor foolproof, either—that is not technically possible. Nonetheless, these tools can provide the necessary "friction" that makes undesirable use of AI that much harder to execute and they will also create "footholds" for the policy action.

## Policy solutions

As I noted above, technological approaches will need to work hand-in-hand with policy. Here are some possible policy approaches to pursue.

**AI-generated content disclosure requirement.** One relatively straightforward step would be to require that any consumer-facing AI-generated content be labeled as such. This kind of mandatory disclosure would, for example, likely hamper an AI-powered persuasion campaign we described above—at least, as long as this rule was abided by.

Of course, deciding the exact level of AI involvement that would trigger such a mandate—as well as the form it would need to take—would require careful deliberation. And the rules would have to be updated as the technology and the use of it evolved. In particular, it would be important to avoid the "user desensitization" effect, in which the users stop paying attention to the corresponding disclosures due to being bombarded with them at every occasion (and for trivial reasons). (Such desensitization seems to have occurred, for example, in the context of the web cookie usage disclosure

and consent requirements imposed in the European General Data Protection Regulation (GDPR) [5].)

**Accelerating the use of content authenticity tools.** As discussed earlier, content authenticity tools such as watermarking, deepfake detection, protection against unauthorized AI-powered editing, or provenance certification can be very useful but their effectiveness is hardly guaranteed. Even leaving aside technical questions, the efficacy of these solutions will critically depend on how broadly adopted they are. We need here a broad cooperation of the industry players that develop the relevant AI systems, so as to establish consistent expectations and standards. Policy can accelerate this process and broaden the use of such techniques, through incentives and/or mandates. After all, we don't know if market incentives will ever be sufficiently strong to drive the development and deployment of these technologies; they certainly are not enough at this point.

**Client identification and suspicious activity reporting mandates.** One possible approach to deterring rogue actors could be adapted from anti-money laundering laws. It would require providers of sufficiently capable AI services to implement adequate client identification mechanisms. These AI providers would then be expected to monitor the usage of the tools they supply to flag (and, potentially, block) suspicious activity as well as to report it to appropriate governmental agencies (such as FBI) or other organizations.

**Advance "AI literacy" efforts.** Of course, no technical solution or set of regulations will ever suffice to fully mitigate the risks AI now poses. It is thus crucial that, in addition to "email literacy" and "social media literacy," we think about promotion of "AI literacy." The public needs to understand how to judiciously interact with AI systems—and how to be on the lookout for when they are interacting with AI in the first place. This includes helping the public avoid the natural tendency to anthropomorphize AI systems. After all, AI does not reason; it merely mimics reasoning—at least as of now. We also must go from assuming that content is authentic until proven otherwise to assuming that content is fake until proven otherwise—or at the very least discounting the value of unverified content.

Overall, there is a need for a shift in the public mindset to accommodate how AI is changing the world. We thus need a decisive policy thinking on how to advance such AI literacy more intentionally, instead of relying on our society learning it the "hard way."

To conclude, let me reiterate that I am excited about the positive impacts that AI can have, but I also want to be clear about and mindful of the risks it gives rise to. Today, my aim is to highlight one family of such risks. I am optimistic that we can mitigate these risks, but this will require work. It cannot be left to chance. And we need to get started now.

Thank you and I am looking forward to your questions.

## Acknowledgements

I am grateful for invaluable help from Sarah Cen, David Goldston, Andrew Ilyas, and Luis Videgaray.

## References


[1] Sai Balasubramanian. AI offers promise and peril in tackling loneliness. *Forbes*. https://www.forbes.com/sites/saibala/2023/05/17/can-artificial-intelligence-solve-the-growing-mental-health-crisis/.

[2] Karen Hao. Deepfake porn is ruining women's lives. Now the law may finally ban it. *Technology Review*. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

[3] Faith Karimi. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. *CNN.com*. https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html.

[4] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *Arxiv preprint arXiv:2306.04634*, 2023.

[5] Oksana Kulyk, Nina Gerber, Annika Hilt, and Melanie Volkamer. Has the GDPR hype affected users' reaction to cookie disclaimers? *Journal of Cybersecurity*, 6(1), 12 2020.

[6] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes. In *Arxiv preprint arXiv:2004.11138*, 2020.

[7] Johan Moreno. Canva opens up access to docs in beta, adds "magic write" generative AI copywriting tools. *Forbes*. https://www.forbes.com/sites/johanmoreno/2022/12/07/canva-opens-up-access-to-docs-in-beta-adds-magic-write-generative-ai-copywriting-tools/.

[8] Vivek H. Murthy. Our epidemic of loneliness and isolation. *The U.S. Surgeon General's Advisory*. https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf.

[9] Donie O'Sullivan and Jon Passantino. 'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion. *CNN.com*. https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html.

[10] Kevin Roose. An A.I.-generated picture won an art prize. artists aren't happy. *New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.

[11] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Mądry. Raising the cost of malicious ai-powered image editing. In *Arxiv preprint arXiv:2302.06588*, 2023.

[12] Pranshu Verma. They thought loved ones were calling for help. It was an AI scam. *The Washington Post*. https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.

[13] James Vincent. DeSantis attack ad uses fake AI images of Trump embracing Fauci. *The Verge*. https://www.theverge.com/2023/6/8/23753626/deepfake-political-attack-ad-ron-desantis-donald-trump-anthony-fauci.

[14] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. In *Arxiv preprint arXiv:2305.20030*, 2023.