Cybersecurity, Information Technology, and Government Innovation Subcommittee

Hearing: "Advances in AI: Are We Ready For a Tech Revolution?"

Written Statement of Aleksander Mądry

March 8th, 2023

Chairwoman Mace, Ranking Member Connolly, and Members of the Committee, thank you for inviting me to testify. Today, I want to share my perspective on the recent advances in AI as well as the key opportunities and challenges they give rise to.

I want to start by stating the obvious: AI is no longer a matter of science fiction nor is it a technology confined to research labs. AI is a technology that is already being deployed and broadly adopted as we speak. It will drastically change our lives; we need to be thinking now how to shape the AI-driven world that comes.

In what follows, I will elaborate on four overarching points:

1. AI systems, particularly those with natural-language interfaces, will become deeply integrated into our economy and society.

2. These AI systems give rise to many promising opportunities but also risks. We must carefully and proactively balance the former with the latter.

3. When it comes to regulating the new generation of AI systems, anthropomorphizing AI can be unhelpful and even misleading.

4. It is critical that we pay attention to the emerging AI supply chain. This chain engenders a number of reliability and regulatory challenges. It will also structure the distribution of power in an AI-driven world.

What makes the latest AI tools so pervasive?

Much of the recent interest in AI has been driven by the family of AI tools known as *generative AI*. ChatGPT[1], DALL·E 2[2], Midjourney[3], Stable Diffusion[4], and Bing Chat[5]—all released in just the past year—rely on generative AI. These tools are leaps and bounds more powerful—and more popular—than those that came before them. Why?

One edge these recent AI tools have over those developed previously is their *ease of use*. In particular, these tools have sophisticated natural language interfaces that are able to process user requests that are given in plain English (and, often, in some other human languages too). Take ChatGPT, for instance—an AI-powered chatbot developed by OpenAI. Users can simply command ChatGPT to "Draft a memo about the upcoming hiring meeting" or "Compose an email to a vendor complaining about the suitcase we bought from them" in much the same way they would ask a human assistant.

The other key differentiating factor for these recently released AI tools is their *scale*. That is, the basic design principles underlying these tools are fairly simple, but they apply these principles to extraordinarily large AI models (in terms of the number of parameters[6]) that are trained on extraordinarily large amounts of data. To illustrate this point, note that the operating principle behind ChatGPT (and more generally, all such *large language models*, or *LLMs*) is very simple: given a passage, produce a word that would likely come next. (Here, "likely" indicates that the word appears after a similar passage in the training data.) So, given the passage, "Once upon a

[1] https://chat.openai.com
[2] https://openai.com/product/dall-e-2
[3] https://midjourney.com
[4] https://stability.ai/
[5] https://www.bing.com/new
[6] One can think of parameters as degrees of freedom—or "knobs"—that the model has to fit the patterns that the training data exemplifies. The number of parameters is thus a natural measure of the model's complexity and its ability to extract more nuanced insights from that training data.

time, in a land far far," the model would suggest that "away" is most likely the next word. Or perhaps more saliently, for a query passage "What's the capital of France?," it would suggest that the likely word to follow is "Paris."

If the underlying principle is so simple—and, in fact, it has been already leveraged in countless AI systems developed in the past—why has an impressive tool like ChatGPT emerged only recently? The answer is, again, *scale*. On the one hand, the size of ChatGPT's training set is so large—and thus contains many example passages to learn from—that it enables ChatGPT to guess likely next words in a way that sometimes feels prescient. To put it in numbers, the training set for a system like ChatGPT is at least 160 times larger than the size of all of the text of Wikipedia.

To train an AI system effectively on such a massive dataset, one also needs the underlying model to be extremely large—of the order of hundreds of billions of parameters. The resulting scale of computational resources (in terms of the hardware and computation time) is mind-boggling. While there are no official sources, some estimates put the development and operation cost for OpenAI's ChatGPT at hundreds of millions of dollars. (And these numbers are likely to climb much higher for the next generation of even more capable systems.)

All in all, the unprecedented amounts of data and computing resources have made recent AI tools possible, and the ease with which people can use these natural-language tools is what will make them more powerful and pervasive than anything that has come before.

Promises of large-scale generative AI

What does the unprecedented scale (and cost) of current AI give us? Tools that have impressive capabilities, and that can be used with ease—no AI expertise required. This tremendous accessibility—together with a broad set of potential uses—is a key factor driving the rapid adoption

of AI.

Indeed, it is not hard to imagine many ways in which AI tools could assist us in mental and creative tasks, bolstering human capabilities and empowering us. For example, ChatGPT is already highly effective at summarizing and synthesizing text, and has inspired startups that automate industries from news summarization to copy editing. Tools like Midjourney (an AI tool for generating hyper-realistic images using a natural-language interface) are transforming the world of visual design and art (while also spurring a number of legal disputes regarding copyright law [5, 10]). GitHub Copilot[7]—an AI tool that generates code guided by natural language prompts—is quickly becoming an indispensable resource for any efficient programmer.

The next generation of such AI tools is poised to deliver even more. They could help us efficiently integrate and leverage our accumulated knowledge—be it in healthcare, science, or engineering—to suggest a promising experiment to run, material to design, or hypothesis to test. They may provide support for individual education at scale by supplying each student with a personalized and infinitely patient tutor. They can create a future in which AI tools help us perform tedious, resource-intensive tasks, making our lives not only more productive but also more fulfilling.

Risks of mass AI adoption

These positive and empowering impacts of AI are possibility but hardly preordained. The adoption of AI comes with a multitude of risks—some that have been long studied in the field, and others that are only now emerging due to the scale of recent AI systems. Mitigating these risks and monitoring them—for example, through auditing—should be of the highest priority. We discuss several such risks below.

Robustness and reliability. AI systems are often brittle—their performance rapidly degrades

[7] https://github.com/features/copilot

when they are faced with data that differs from what they have encountered in their training data [17, 18]. This brittleness manifests in a variety of ways, even in older (more well-studied) AI systems. In particular, there is large body of academic work both studying *distribution shift* [17, 18]—in which AI systems over-rely on their training data and thus degrade in performance when acting on less familiar data—and *adversarial vulnerability* [2, 16, 1]—in which a malicious actor can manipulate the query to an AI model in order to get a desired output. It turns out that these issues continue to arise in the context of the latest generation of AI tools, for example, in the context of large language models (such as ChatGPT):

- AI systems continue to over-rely on the training data. On the one hand, similarly to the distribution shift problem, this over-reliance can lead to "hallucinations:" totally inaccurate answers to unfamiliar questions (i.e., questions not found in the training data). On the other hand, this over-reliance also leads to mistakes on data that share surface-level resemblance to training data—for example, when faced with the question "What weighs more: a pound of feathers or two pounds of gold," ChatGPT responds that they "both weigh the same amount," likely due to the question's similarity to a classic riddle appearing many times in its training data.

- At the same time, large-scale AI systems remain vulnerable to adversarial manipulation. Largely innocuously, for example, users have elicited professions of love, threats of harm, and inappropriate content from them using a technique known as "prompt engineering."

- Worse still is that even AI experts have a very poor understanding of the extent of these issues, and are only beginning to find ways to discover and alleviate them.

Perpetuating data and designer biases. The data used to train modern AI systems is vast

and deeply flawed. For example, large language models are often trained on data from Reddit, which contains a myriad of offensive and inappropriate content. Research shows that AI systems can perpetuate, and even promote the biases present in this data [17]. What's more, these systems can also propagate biases held (even unknowingly) by the system designers themselves.

Facilitating fake media, phishing, and harassment. The accessibility of modern AI systems, combined with their ability to rapidly produce cogent text makes them an invaluable tool for scammers, trolls, and other malicious actors. Even in their infancy, large-scale AI systems have already been used to defraud innocent people [19]; to generate intimate images of ex-partners (so-called "deep-fake revenge pornography" [12]; and to create and massively disseminate disinformation [20]. As AI systems become more powerful (and thus more believable) and even easier to use, the risks of misuse will only increase [7]—and while technical solutions can help here, the question is ultimately one of policy [4, 15, 14].

Still, although these risks are very real, they are also manageable. As we've shown, many of these risks have already been identified and studied. With continued effort, we can put ourselves in a good position to control rather than be controlled by the deployment of AI.

Human intuition can be a liability

As we engage with AI more directly, we expose ourselves to interactions that counter our intuitions and exploit our cognitive biases. Indeed, perhaps because we're now able to communicate with AI tools so seamlessly, it's easy for us to think of them as human. But this is a mistake. These tools are not human. They are simply pattern extractors—a simple principle applied at massive scale [9]. By anthropomorphizing AI tools, we fool ourselves into thinking that we understand how they behave—or that we can straightforwardly adapt policies designed for humans to work in AI

contexts. Both of these intuitions are incorrect, and can be dangerous.

ChatGPT, for example, often generates completely nonsensical or incorrect answers[8]. These answers, however, are phrased so convincingly, so much *like* a human, that we tend either to (a) believe these answers outright; or (b) believe that we understand the "reasoning" that went into the answer[9]. *In reality, however, even AI experts do not understand why exactly ChatGPT generates what it does.* And while we expect AI systems to produce incorrect answers less frequently as they improve and evolve, having blind faith in these models will result in unexpected failures and can (more worryingly) erode our analytical and sense-making abilities. We cannot just depend on AI getting better—helping users validate and verify the outputs of large-scale AI systems should be a key policy, technological, and educational goal. From a technical and policy standpoint, we also need to rethink what sorts of *guarantees* (if any) these systems can provide on their outputs—I return to this issue in the next section.

For the same reason, approaching AI policy as we would previous, human-driven technologies (or humans themselves) is bound to yield blind spots. For example, there are sometimes legal requirements under which decision-makers must justify or explain their decision to the decision subject. If the decision maker is a human, we might simply ask them for a plain-text explanation. If these decisions are made by an AI system with a conversational interface (such as ChatGPT), one might be tempted to apply the same requirement—to simply ask the AI tool to explain its reasoning. However, this approach would be misguided. We cannot straightforwardly adapt every existing policy to AI systems. What is needed instead is a robust technical and legal framework

[8]For example, asking (current version of) ChatGPT "how many countries start with the letter V" yields a confident "There are no countries whose names start with the letter V." When asked to confirm, ChatGPT responds "Yes, I'm sure. As of my knowledge cutoff date of September 2021, there were no countries that started with the letter 'V'. It's possible that a new country may have been established since then or..."

[9]"Reasoning" here is in quotation marks since there is substantial disagreement, even at the forefront of AI research, about whether what ChatGPT does can or should be called "reasoning."

for handling decisions made by AI systems. Developing such a framework is an active area of both technical and legal research [21, 13, 8].

A key factor shaping AI's impact: The AI supply chain

AI technology will reshape our society not only through our direct interaction with it as individuals, but also by the way it will be developed and used by companies in almost all sectors of the economy.

A key feature of emerging large AI systems is that, while consumers can interact with them (as with ChatGPT), they can also be used as a base (or foundation [3]) on which to build other systems, effectively becoming the "upstream" link in a layered *AI supply chain*.

For example, consider a company using AI in its hiring process. It's unlikely that the company would build its own AI hiring system from scratch. Instead, the company would likely combine a few hiring AI models, each developed by a third-party company. These third-party models might leverage AI tools that were developed by yet other third-party companies. A typical AI supply chain is usually many layers deep. And at the upstream of this chain might be a base model like GPT [6] (the engine that powers ChatGPT).

Such a structure makes perfect economic sense. Base models are extraordinarily expensive to develop. Creating such models is thus out of reach for a company that just wants to use AI for hiring, and even out of reach for all its third-party AI providers. Instead, those companies will build on top of a base model, enabling the tools in this chain to leverage the base model's advanced capabilities to increase performance on their specific, desired tasks (e.g., parsing resumes, or summarizing job interviews). As a result, base models will be few, but one should expect an almost Cambrian explosion of startups, new use cases and business models in the "downstream" of the AI supply chain—tools that all make use of the base models. We can see the first glimpses of this

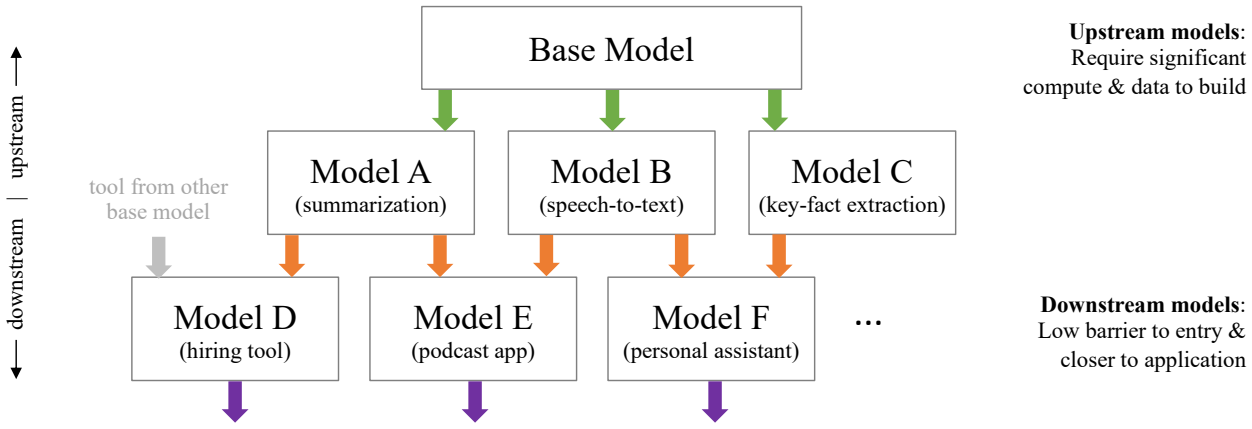

Figure 1: *An example of the AI supply chain.* AI models will be built on top of one another, forming a layered AI supply chain. At the top of the supply chain (upstream) are base models that require significant computation and data to build. Further down the supply chain (downstream) are AI tools that are built on top of other AI tools. Downstream models are generally easier to build and use than upstream ones.

already, with over 900 companies/organizations[10] worldwide already using OpenAI base models in their business processes in sectors as diverse as education, manufacturing, finance, retail, healthcare and government. (In fact, just in a single week last month, Coca Cola[11], Spotify[12] and Snap[13] announced their plans to incorporate AI that builds on top of OpenAI's systems into their business processes.)

With their ease of use and powerful capabilities, upstream AI base models are poised to permeate a sizable fraction of economic activity. This begs the question: who will develop and provide these systems? The likely answer is that, given the skills and enormous capital investment that building such systems requires, only very few companies will be able to compete here.

This leads to a couple of policy-relevant observations. First, the reliance of the AI supply chains on a limited number of large upstream systems introduces potential vulnerabilities across the many

[10] https://tinyurl.com/bdem9uhy
[11] https://tinyurl.com/29332cnu
[12] https://newsroom.spotify.com/2023-02-22/spotify-debuts-a-new-ai-dj-right-in-your-pocket/
[13] https://techcrunch.com/2023/02/27/snapchat-launches-an-ai-chatbot-powered-by-openais-gpt-technology/

users of these systems. Structural biases or hidden systemic fragilities could have impacts across the entire economy. Suppose the hiring AI tool in our example has some bias or deficiency stemming from a large base model—whether because of an error in data collection or an issue in that tool's training or some unexamined biases of its creators. Let's say this bias or deficiency led the system to be skeptical about applicants graduating from a particular high school. In that case, every downstream system that used this tool would likely inherit this bias or deficiency, making it systematically harder for graduates from that high school to find *any* job. Or, if a large number of financial institutions relied on the same large language model for analysis related to asset-management, the financial system might be prone to herd-behavior and hidden fragilities [11]. Also, imagine if one of the upstream models went suddenly offline—what would happen downstream?

If one of the emerging risks is that we may become too dependent on a very few upstream models, the flip side also raises problems. As already noted, there will likely be a proliferation of companies building "downstream" applications that depend on these base models. AI users will be dealing with systems that are a product of the interactions of layers of AI systems.

The properties of such "combined" or "composite" systems might become even more unpredictable and idiosyncratic. Consider our hiring example above. Suppose that the company's (AI-powered) hiring policy is found to be discriminatory by the U.S. Equal Employment Opportunity Commission (EEOC). Who should shoulder the blame and the responsibility for fixing the problem? The employer? The company that sold the AI hiring tool? The curator of the data driving this tool? Or the developer of the large upstream model? Should each entity along the AI supply chain be required to meet a set of specifications or communicate their design process? Policy makers will have to grapple with these kinds of questions, and likely will need to do so soon. Otherwise, this growing AI supply chain complexity may result in situations where the systems do not perform as

expected despite everyone involved fulfilling their duty of care.

As noted above, this is an even more complex problem because AI experts cannot explain how individual AI systems reach their conclusions, and figuring that out becomes much more difficult when systems are layered on top of each other.

All these challenges draw attention to a fact that is rarely acknowledged in academic and policy circles: current AI technology is not particularly well suited for deployment through complex supply chains. As opposed to conventional software, which works well when deployed through many layers (the "software stack"), *AI technology has fundamental attributes that make layered deployment problematic.*

First, AI systems tend to interact with each other in a non-modular way. That is, connecting two AI systems to each other may change their individual attributes in unpredictable ways. Second, there can be hidden data overlaps or correlations that can prevent a composite system from performing as expected. And third, AI systems—including the very large ones—do not provide performance guarantees (e.g., guarantees that come with reliable validation and auditing procedures). Consequently, regulatory and policy initiatives that overlook the impact of complexity in the AI supply chain seem destined for failure.

Finally, we should not forget that this shaping up of the AI supply chain is going to structure power—the control over where, when, and how AI is used. We thus need to pay attention not only to the nature of AI as a technology, but also to the way it is deployed and by whom. These factors will be paramount—from a societal standpoint, from a geopolitical standpoint, and from a national security standpoint.

To conclude, let me go back to the beginning of my testimony and say: we are at an inflection point in terms of what future AI will bring. Seizing this opportunity and bringing about a future

that is positive and empowering requires discussing the role of AI in our society and nation, what we want AI to do (and not do) for us, and how we ensure that it benefits us all. This is bound to be a difficult conversation, but we do *need* to have it, and have it *now*.

Acknowledgements

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

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases (ECML-KDD)*, 2013.

[3] Rishi Bommasani et al. On the opportunities and risks of foundation models. In *Arxiv preprint arXiv:2108.07258*, 2021.

[4] Dan Boneh, Andrew J. Grott, Patrick McDaniel, and Nicolas Papernot. Preparing for the age of deepfakes and disinformation. *HAI Policy Brief*. https://hai.stanford.edu/policy-brief-preparing-age-deepfakes-and-disinformation.

[5] Blake Brittain. AI-created images lose U.S. copyrights in test for new technol-

ogy. *Reuters*. https://www.reuters.com/legal/ai-created-images-lose-us-copyrights-test-new-technology-2023-02-22/.

[6] Tom B. Brown et al. Language models are few-shot learners. In *Arxiv preprint arXiv:2005.14165*, 2020.

[7] Miles Brundage et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. In *Arxiv preprint arXiv:1802.07228*, 2019.

[8] Sarah H. Cen and Manish Raghavan. The right to be an exception to a data-driven rule. In *Arxiv preprint arXiv:2212.13995*, 2022.

[9] Ted Chiang. Chatgpt is a blurry jpeg of the web. *The New Yorker*. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

[10] Nicole Clark. Artists sue AI art generators over copyright infringement. *Polygon*. https://www.polygon.com/23558946/ai-art-lawsuit-stability-stable-diffusion-deviantart-midjourney.

[11] Gary Gensler and Lily Bailey. Deep learning and financial stability. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3723132.

[12] Douglas A. Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke law and technology review*, 17:99–127, 2019.

[13] Margot E. Kaminski and Jennifer M. Urban. The right to contest ai. *Columbia Business Law Review*, 7:1957–2048, 2021.

[14] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Arxiv preprint arXiv:2301.10226*, 2023.

[15] 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.

[16] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[17] Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

[18] Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. A taxonomy and terminology of adversarial machine learning. https://csrc.nist.gov/publications/detail/nistir/8269/draft.

[19] 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/.

[20] John Villasenor. How to deal with ai-enabled disinformation. *Brookings*. https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation.

[21] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Columbia Business Law Review*, 2:494–620, 2019.