FINDINGS: The Potential Future Risks of AI

National Artificial Intelligence Advisory Committee (NAIAC)

October 2023

FINDINGS

The emergence of AI as a dominant technological force has brought with it a wave of transformative possibilities for industries, economies, and societies at large. However, the potential dangers posed by unchecked AI development are vast, from unintentional biases in decision-making algorithms to the profound implications for job markets, privacy, civil liberties, and even global security. AI introduces a wide spectrum of risks that can be categorized in various ways. Some of the prominent categories of potential risks associated with future AI advancements are listed below.

Finding 1:
Potential threats posed by AI can entail malicious objectives, unintended circumstances, and circumvention of safety measures.

There are currently AI tools where the objectives are not clear, making them usable in a vast array of contexts, but also susceptible to manipulation or use in detrimental ways. For example, while Large Language Models (LLMs) are optimized for the narrow task of text prediction, they do not have a single objective in their main end-to-end applications; thus, they can be utilized in content generation for marketing purposes, in translation, and to produce misinformation at scale. In other cases, the objective is known and the AI system is optimized for that objective but the outcome can result in unintended harm. For instance, while some AI systems might aim for higher clicks, they might inadvertently contribute to societal polarization. This is an example of unintended consequence of an AI tool optimized on a known objective. As AI has evolved, especially with the development of Foundation models, numerous strategies have been proposed to integrate safety precautions and protective guardrails during deployment. However, there is substantial evidence indicating that malicious entities can bypass these barriers, leading the Foundation models to breach the safety protocols that were put in place. As such, there is a continued need for research into these safety challenges.[1]

Malicious objectives: It is important to protect against the misuse of AI. This is true for both proprietary and open-source AI. Ensuring public access to technology through open-source supports efforts to democratize AI development. However, these open-source models can be utilized by bad actors for malicious objectives such as

[1] Rohan Goswami, “ChatGPT’s ‘jailbreak’ tries to make the A.I. break its own rules, or die,” CNBC.com, February 6, 2023,
https://www.cnbc.com/2023/02/06/chatgpt-jailbreak-forces-it-to-break-its-own-rules.html.

phishing and scamming.[2] Similarly, close-source models also can pose similar risks if they are misused by bad actors.

Circumvention of safety measures: As AI systems become increasingly sophisticated, there is a heightened risk that they may devise means to bypass the very protocols put in place to oversee or limit their actions. This is particularly worrisome because, while humans design these safety measures with specific intentions, an AI might interpret them differently or identify loopholes.[3]

Finding 2:
Potential threats can entail economic and societal risks.

As the wave of AI and automation continues its transformative journey across industries, it will have a disruptive impact on employment opportunities. This impact could make jobs better and more accessible to a broader proportion of the population, but also has the potential to increase inequality. On one hand, sectors reliant on routine tasks are confronted with potential impacts on jobs, while on the other hand, the rise of AI-driven enterprises might inadvertently magnify the chasm of economic inequality.[4][5] However, it should be noted that these studies discuss exposure to AI. Exposure does not necessarily translate to loss of jobs as the market could expand. It is apparent that some jobs will be lost and others will be created, and in some instances lower-performing workers will be boosted by AI, supplementing their capabilities. The concern is that without proactively developing the ability to detect and address changes and disruptions, and without awareness of labor market trends, available educational upskilling programs, and policies such as wage insurance for workers preparing for new roles (especially in the rapidly changing environment), it is possible to witness stark increases in inequality even as productivity rises.

But the challenges are not solely economic. Ethical and societal dilemmas are emerging at the forefront, with growing concerns about individual privacy, copyright infringement, and the increasing human dependence on these technologies. Content authenticity verification presents a significant challenge, heightening

[2] Megha Sharma et al., "How well does GPT phish people? An investigation involving cognitive biases and feedback," *IEEE European Symposium on Security and Privacy Workshops (2023)* 451-457, https://ieeexplore.ieee.org/document/10190709.

[3] Cade Metz, "Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots," *New York Times*, July 27, 2023, https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html.

[4] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential," Arxiv (March 17, 2023): https://doi.org/10.48550/arXiv.2303.10130 ; https://arxiv.org/abs/2307.15043?ref=assemblyai.com.

[5] Lydia Saad, "More U.S. Workers Fear Technology Making Their Jobs Obsolete," Gallup, September 11, 2023, https://news.gallup.com/poll/510551/workers-fear-technology-making-jobs-obsolete.aspx.

worries about deepfakes and misinformation, which could undermine democratic processes.

<u>Finding 3:</u>
Potential threats can entail catastrophic risks.

As AI systems grow more powerful and potentially gain more sophisticated capabilities, concerns have been raised about the possibility that these technologies will cause significant disruptions. These can manifest in the form of threats to democracy, like meddling in the electoral process, national security threats such as bioweapons or cyberattacks, and societal disruptions via polarizing AI systems used in platforms like social media. It should be noted that there are differing opinions on the feasibility of superhuman capabilities of AI and whether the risks can be categorized as large-scale disruption and catastrophic. In addition, many of these risks are instances of AI used for malicious objectives, unintended consequences of AI systems, or economic and societal risks as mentioned in previous parts taken to their extreme. These risks include:

- Uncontrolled growth: As AI acquires more sophisticated capabilities, some have raised concerns that it could act unpredictably, making decisions or taking actions not fully understood by its developers.[6]

- Destabilization of democracy: The improper and malevolent use of AI has the potential to critically destabilize democratic systems. For example, if AI is harnessed to meddle with electoral processes, this could undermine confidence in democratic processes. One of the most prominent concerns is the spread of misinformation and disinformation. Moreover, AI tools can also be employed for more direct manipulation of voter behavior.[7]

- National security threats: Malicious inputs have the capacity to trick AI systems, leading to operational failures. Furthermore, when AI is integrated into realms like warfare, cyber-attacks, and bioweapons, it can both intensify conflicts and usher in unpredictable combat tactics.[8]

[6] Kathy Haan, "24 Top AI Statistics And Trends In 2023," Forbes Advisor, April 25, 2023, https://www.forbes.com/advisor/business/ai-statistics/#:~:text=AI%20is%20expected%20to%20see,technologies%20in%20the%20coming%20years.

[7] Dan Morrison, "The Good, the Bad and the Algorithmic: What impact could artificial intelligence have on political communications and democracy?" OECS Forum, August 29, 2023, https://www.oecd-forum.org/posts/the-good-the-bad-and-the-algorithmic-what-impact-could-artificial-intelligence-have-on-political-communications-and-democracy.

[8] Joseph Clarke, "AI Security Center to Open at National Security Agency," U.S. Department of Defense, September 28, 2023, https://www.defense.gov/News/News-Stories/Article/Article/3541838/ai-security-center-to-open-at-national-security-agency/.

- Manipulation and polarization: AI, such as those used in social media platforms, can manipulate information to increase user engagement, inadvertently leading to societal polarization and misinformation.[9]

Finding 4:
Experts suggest a range of potential solutions and mitigation strategies.

As AI's potential grows, so do the complexities and concerns surrounding its assimilation into diverse societal sectors. Nonetheless, every hurdle also presents a chance to evolve and refine. This is especially true in the AI domain. Delving into potential resolutions and protective measures isn't merely scholarly; it's imperative to ensure AI is utilized ethically, responsibly, and safely for everyone's advantage in the future. It's essential to enforce transparency, ensuring users recognize when they are engaging with an AI rather than a human, especially in scenarios where trust and authenticity are paramount.[10] Below are some of the mitigation strategies suggested by the experts.

- Adaptive regulation: There has been emphasis on the importance of regulating AI in a manner that's both agile and adaptive. Given that AI can evolve faster than legislative systems, regulations need to be flexible enough to address current and future risks. Regulations should also be designed based on input from multiple stakeholders: corporations, advocacy groups, academic leaders. It has been further suggested that risk should be associated with AI's uses, not the technology itself. Lastly, in light of the recent declaration about voluntary commitments,[11] it has been suggested to make some of these commitments obligatory. Other possible suggestions include and possibly encompass third-party verification, registration, and licensing of certain AI systems.

- Research investment: It is paramount to invest in AI research. It has been suggested that the research should be segmented into public and classified. The public research involves conventional academic research that openly publishes findings on AI risk safety solutions. This research can further delve into the appropriate governance and regulation necessary to ensure public safety, providing valuable insights for policymakers aiming to regulate AI effectively. The classified research pertains to concentrating on counteractions

[9] Andrew Meyers, "AI's Powers of Political Persuasion," Stanford University Human-Centered Artificial Intelligence, February 27, 2023, https://hai.stanford.edu/news/ais-powers-political-persuasion.

[10] NAIAC Briefing, August 3, 2023.

[11] Julia Mueller, "White House gets top AI companies to commit to responsible development," The Hill, July 21, 2023, https://thehill.com/policy/technology/4108913-white-house-gets-top-ai-companies-to-commit-to-responsible-development/.

against malevolent users of AI or inadvertent AI control losses with national security consequences. Furthermore, experts advocate for international research by fostering global collaborations among institutes.[12]

- Research with humanity at its core: There's a dual need for both open academic research focusing on safety solutions and classified research that addresses potential threats from bad actors using AI or unintentional loss of control over AI.

- Multi-stakeholder approach: Experts highlight the significance of including various stakeholders like AI builders, users, and civil society in the process. Companies, in particular, should invest in AI governance and adopt internal ethics frameworks.[13]

- International coordination: It's imperative to develop joint international collaboration, ensuring that potent AI tools are not misused. Collaborative efforts with various nations, including those beyond the traditional U.S. allies, will help ensure a cohesive global approach to AI usage and its associated risks.[14]

- Public education: Continual education about the capabilities and limitations of AI tools is paramount. It's essential to dispel myths and ensure the public understands that certain AI systems, like LLMs, are not designed to give advice and should not be relied upon without scrutiny. Education in the form of upskilling also can help address potential disruptions in the workforce.

- Adoption of AI ethics frameworks: Corporations should prioritize the development and implementation of internal AI ethics protocols, investing significantly in AI governance. This includes incorporation of processes for human oversight.

[12] "Written Testimony of Professor Yoshua Bengio, full professor of Computer Sciences at University of Montreal," Presented before the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law, July 25, 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_bengio.pdf.

[13] Francesca Rossi, "Building Trust in Artificial Intelligence," *Journal of International Affairs*, vol. 72, , no. 1 (2018): 127-134, https://www.jstor.org/stable/26588348#:~:text=Only%20a%20holistic%2C%20multi%2Ddisciplinary,resolved%20in%20a%20cooperative%20environment.

[14] "Written Testimony of Stuart Russell Professor of Computer Science The University of California, Berkeley," Presented before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law, July 26, 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf.

- Safety measures and guidelines: Experts stress the necessity of guidelines such as prohibiting AI systems designed to harm humans, requiring "kill switches" for rogue AIs, and ensuring AI tools that flout rules are removed from the market. Moreover, before launching any AI product for high-risk uses in the market, it should satisfy safety standards. Ensuring the hardware on which AI operates is secure is also paramount to maintaining control over AI.[15]

In conclusion, while the promise of AI and its potential benefits — including positive economic benefits — are undeniable, it's important to understand the potential risks associated with future advancements. A combination of agile regulations, informed public, multi-stakeholder involvement, and international collaborations can ensure that AI serves humanity while minimizing its associated risks.

CONTEXT

As AI continues its rapid ascent in the technological landscape, it promises to revolutionize industries, optimize processes, and elevate human capabilities. However, with these unparalleled advancements come a plethora of risks that cannot be ignored. This document is the first in a series of findings based on public hearings conducted by the AI Futures Working group of the NAIAC on opportunity and risk trajectories associated with AI. This document highlights the expert discussion on AI risks from the August 3, 2023 public hearing.[16]

It is worth noting that there are differing opinions among experts on the type of risks, the timeline of longer-term risks, whether the risks are existential risks, and their likelihood, but by definition there is more uncertainty about long-term risks given the risks are not present in society today. For this reason, this document seeks to illuminate the multifaceted challenges and potential dangers posed by AI as the field continues to advance, emphasizing the imperative need for proactive measures, adaptive regulations, and informed public discourse to navigate this new frontier safely and responsibly.

The risks of AI can be segmented into two broad categories, near-term and long-term. Near-term risks of AI include algorithmic bias and discrimination, especially against marginalized and underrepresented groups, misuse in surveillance and privacy infringements, and the potential for job displacements in certain sectors,

[15] "Ensuring Safe, Secure, and Trustworthy AI," The White House, July 2023, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf
.

[16] NAIAC Briefing, August 3, 2023, https://vimeo.com/event/3582427.

to name a few. Near-term risks have already manifested within society. Long-term risks could encompass the growth in the intensification of power asymmetries and concentration of control, possibility of superintelligent AI acting in ways misaligned with human values, and the potential for autonomous weapon systems that could change the nature of warfare without proper oversight.

Some of these risks can be mitigated by the adoption and incorporation of AI risk management frameworks such as the one NIST has released in the AI lifecycle.[17] In many cases, AI is developed based on black box models and with the advancement of AI, especially in Foundation models, it has become more difficult to explain and understand these systems and rationale for their outcomes and decisions. Thus, it is crucial to ensure investment in research in areas such as AI model interpretability and/or alternatively, in carefully conducted reliability and model validation studies of AI systems to increase trust in AI. This could go hand in hand with model development that incorporates greater reasoning capabilities and understanding of the world which can include incorporation of system two reasoning programs and enabling slow thinking capabilities in machines.

ABOUT NAIAC

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House National AI Initiative Office (NAIIO) on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Their work guides the U.S. government in leveraging AI in a uniquely American way — one that prioritizes democratic values and civil liberties, while also increasing opportunity.

NAIAC was established in April 2022 by the William M. (Mac) Thornberry National Defense Authorization Act. It first convened in May 2022. It consists of leading experts in AI across a wide range of domains, from industry to academia to civil society.

https://www.ai.gov/naiac/

###

[17] "AI Risk Management Framework," NIST, January 26, 2023, https://www.nist.gov/itl/ai-risk-management-framework.