Testimony

Advancing Trustworthy Artificial Intelligence

Addendum

Jason Matheny

CT-A2824-2

Document submitted August 1, 2023, as an addendum to testimony before the U.S. House Committee on Science, Space, and Technology on June 22, 2023

For more information on this publication, visit www.rand.org/t/CTA2824-2.

Published by the RAND Corporation, Santa Monica, Calif.

Advancing Trustworthy Artificial Intelligence

Testimony of Jason Matheny[1]
The RAND Corporation[2]

Addendum to testimony before the Committee on Science, Space, and Technology
United States House of Representatives

Submitted August 1, 2023

Following the hearing on June 22, 2023, the congressional committee sought additional information and requested answers to the questions in this document. The answers were submitted for the record.

Question from Chairman Frank Lucas

Question

In your testimony, you suggested one way to prevent AI risks from advanced systems is to place safeguards on entities that use more than a certain threshold of compute. However, the resources needed to deploy models advanced enough to cause significant social harm is dropping significantly. MPT-7B is a recent open-source model that contained 7 billion parameters, was trained in 9.5 days, and cost only $200,000 — far cheaper than previous models of that size.[3]

[1] The opinions and conclusions expressed in this addendum are the author's alone and should not be interpreted as representing those of the RAND Corporation or any of the sponsors of its research.

[2] The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. RAND's mission is enabled through its core values of quality and objectivity and its commitment to integrity and ethical behavior. RAND subjects its research publications to a robust and exacting quality-assurance process; avoids financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursues transparency through the open publication of research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. This testimony is not a research publication, but witnesses affiliated with RAND routinely draw on relevant research conducted in the organization.

[3] All questions are presented verbatim as they were submitted to RAND.

a. Do you think monitoring compute usage will be an effective proxy for model capability as the field continues to innovate?

Answer

This is an important challenge. As more-efficient algorithms and chips for training models are developed, the computational costs of training any given model drop.

Still, compute usage will be an important proxy for model capability. First, even as compute costs drop, the most novel capabilities and risks will likely remain at the frontier of large compute usage, warranting continued oversight of the most compute-intensive models. Second, even if dangerous models eventually become so computationally cheap that oversight is infeasible, monitoring them in their earlier stages will offer crucial opportunities to identify risks and develop countermeasures. Still, compute thresholds can and should be adjusted downward over time, to account for increases in algorithmic efficiency changing what levels of compute pose severe security threats. These adjustments can be done by systematically measuring how much compute is needed to achieve a given artificial intelligence (AI) capability over time, as has been done in some domains.[4]

b. What are the national security implications of model creation costs falling so quickly?

Answer

The decline in model training costs implies that, while limitations on export licenses for powerful chips are likely to remain important, eventually they will not be sufficient. In the absence of further countermeasures, we could see the wide proliferation of models that enable bioterrorism, cyberattacks, large-scale disinformation, and ubiquitous surveillance. To mitigate these threats, it will be important to identify and develop countermeasures for when dangerous models become computationally cheap. These countermeasures could include strengthening biosecurity, improving cyber defenses, verifying which social media accounts are run by humans, developing and disseminating software tools that help AI developers prevent harmful AI accidents, and designing chips that are inherently more secure.

This is not to suggest that there is no role for safeguards focused on computationally expensive models. To the contrary, oversight of computationally expensive models provides early warnings about which threats may be posed by computationally cheap models in the future. Additionally, if computationally expensive models are developed safely, these models may themselves help counter threats from computationally cheap models, such as by improving cyber defenses or flagging bots that are impersonating humans.

A further implication is that, if an innovation would enable the cheap training of models that pose severe security threats, it may be appropriate to delay publication of that innovation until adequate countermeasures have been deployed.

[4] Ege Erdil and Tamay Besiroglu, "Algorithmic Progress in Computer Vision," arXiv:2212.05153 [cs.CV], last updated January 15, 2023.

Question from Ranking Member Zoe Lofgren

Question

I would like to turn the discussion to the government's use of AI, specifically related to use in national security applications. I can understand the argument for developing capability on AI internal to the federal government for security sensitive applications. A security breach could have such dire consequences.

a. Do you think our federal agencies have the resources needed to develop and maintain internal programs and expertise on AI? If no, what can Congress do to help federal agencies gain access to AI resources and talent?

Answer

Federal agencies do not currently have the institutional capacity to build or maintain sufficient internal expertise on AI. The hiring process is slow and cumbersome. A RAND report on hiring for the Department of the Air Force notes that inflexible Office of Personnel Management requirements limited both pay and who managers were able to hire.[5] This research suggests that, in high-skill technology jobs in emerging areas, the federal government is not providing enough flexibility for managers.

Federal agencies are not able to pay market rates for AI experts: The starting salary for someone with an M.S. in machine learning who is working in the San Francisco Bay Area is more than a GS-15 can make. It is unlikely that federal pay will be able to match private sector rates, but RAND research has found that large numbers of private sector experts are willing to make personal sacrifices to serve their country, when given the opportunity.[6] Unfortunately, federal hiring often makes it very difficult for people to serve with long hiring timelines and onerous security clearance processes. The U.S. government can expand the opportunities for Intergovernmental Personnel Act agreements and the use of Schedule A(r) appointments to bring private sector expertise into the government more quickly, and expand the number of part-time and term-limited positions for technology experts to contribute their skills more flexibly. Because many people with relevant talent were born outside the United States, we could also expand service-for-citizenship opportunities, doubling down on one of the United States' greatest advantages: its ability to attract foreign scientists and engineers.

b. The Department of Energy recently released a report which synthesized a series of workshops held in the summer of 2022 to gather input on new and rapidly emerging

[5] Kirsten M. Keller, Maria C. Lytell, and Shreyas Bharadwaj, *Personnel Needs for Department of the Air Force Digital Talent: A Case Study of Software Factories*, RAND Corporation, RR-A550-1, 2022, https://www.rand.org/pubs/research_reports/RRA550-1.html.

[6] Martin C. Libicki, David Senty, and Julia Pollak, *Hackers Wanted: An Examination of the Cybersecurity Labor Market*, RAND Corporation, RR-430, 2014, https://www.rand.org/pubs/research_reports/RR430.html.

opportunities and challenges of scientific AI. The report makes the case for utilizing the long-standing expertise within DOE, especially within the national lab complex. Experts at several DOE laboratories have also used the findings of this report to strongly recommend that the Department's capabilities be leveraged specifically for national security applications. What role do you see the Department of Energy playing as part of the federal government's broad AI initiative?

Answer

The Department of Energy has several strengths that are relevant for AI:

1. large research and development operations, including on AI
2. experts in high-performance computing, which is the basis of current broadly capable AI
3. roles in export controls
4. responsibility for Restricted Data on nuclear weapons secrets, including declassification, which is relevant as soon as AI models can answer sensitive questions about nuclear weapon design
5. a combination of civilian and military concerns and expertise.

That expertise should be brought to bear in handling national security and public safety threats stemming from AI, including developing solutions.

c. Do you think it could make sense for the federal government to utilize private AI systems for national security applications? And if so, can you discuss how we can ensure the security of these systems?

Answer

Yes, through normal administrative processes. The private sector is well ahead of the Department of Defense (DoD) in AI, as intended by the 2022 National Defense Strategy, which plans for the department to be a "fast follower" in the area. The integration of private sector developments into national security applications has long been a strength of the American system.

Existing policies and procedures for cybersecurity work well within government, and efforts toward supply chain security are ongoing. The federal government will need to carefully apply security principles to AI systems, verifying that models are trained on the data they say that they are trained on,[7] that they perform acceptably on a wide variety of intended and unintended inputs, and that, even if they were to severely malfunction, they would not be able to cause harm except through human authorization. Requirements should include restrictions on the ability of models to directly run code that could harm systems vital to national security.

[7] Dami Choi, Yonadav Shavit, and David Duvenaud, "Tools for Verifying Neural Models' Training Data," arXiv:2307.00682 [cs.LG], July 2, 2023.

Questions from Rep. Rich McCormick

Question 1

There are many potential concerns and benefits when it comes to the use of artificial intelligence for our national defense. One contentious matter, which has been popularized in countless science-fiction films, is the use of AI within or in conjunction with weapons systems, especially related to weapons of mass destruction.

a. In your opinion, what standards and limits should Congress and the Defense Community place on the use of artificial intelligence in our military and national security framework?

Answer

DoD has already made substantial steps toward the safe and ethical use of AI through DoD Directive 3000.09 and the Responsible AI Strategy.[8] To prevent technical accidents, unintended escalation, and bias, I think that AI should never be able to make nuclear launch decisions, self-replicate in the style of a botnet, or make personnel decisions that service members do not have a right to appeal. It should also not be able to initiate or escalate a kinetic conflict without rigorous protocols and safety and security testing.

b. At the current state of the technology, is it ideal to utilize AI to perform assistance tasks such as aircraft combat maneuvers in manned or unmanned platforms?

Answer

Before AI can be effectively deployed in a combat environment, it needs not only technical maturity but experienced users who understand and trust the technology that they are working with, as emphasized in the Responsible AI Strategy. There will not be a single moment in which DoD authorizes the use of AI in combat. Rather, existing autopilot and targeting software will improve and be further understood by the warfighters who work with it. While the underlying technology may be good enough to pilot drones under combat conditions, the testing and evaluation frameworks are often not mature enough, or capable of rapid-enough iteration, that they can be effectively used by warfighters or commanders. Prior RAND work has examined some of the existing limitations, which are significant.[9] I recommend that the Department of the

[8] Department of Defense Directive 3000.09, *Autonomy in Weapon Systems*, U.S. Department of Defense, January 25, 2023; U.S. Department of Defense, *U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway*, June 2022.

[9] Danielle C. Tarraf, William Shelton, Edward Parker, Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Léveillé, Jared Mondschein, James Ryseff, Ali Wyne, Dan Elinoff, Edward Geist, Benjamin N. Harris, Eric Hui, Cedric Kenney, Sydne Newberry, Chandler Sachs, Peter Schirmer, Danielle Schlang, Victoria Smith, Abbie Tingstad, Padmaja Vedula, and Kristin Warren, *The Department of Defense Posture for Artificial Intelligence: Assessment and Recommendations*, RAND Corporation, RR-4229-OSD, 2019, https://www.rand.org/pubs/research_reports/RR4229.html; Li Ang Zhang, Jia Xu, Dara Gold, Jeff Hagen, Ajay K.

Air Force focus on building a practice for gradual deployment of AI, on pace with the abilities of these systems.

Question 2

Deep fake imagery and audio, especially in Intelligence and International Influence Campaigns, is a concerning and eerily accurate development with AI. Fake imagery and audio have the potential to create chaos and dangerous misinformation. This capability can easily be used by bad actors to influence elections or the business world, or to blackmail high-level officials.

a. What industry standards or – if necessary – government regulations are appropriate to establish to make sure AI isn't used to cause harm, but to maintain the investment and advancement of this important technology?

Answer

Any viable solutions have to address multiple uses of AI to ensure that sufficiently powerful models are unlikely to cause the greatest harms. While industry standards can serve as a useful initial step, government regulation will be necessary for various domains. Many of those domains are already covered by existing government oversight mechanisms (e.g., self-driving cars are covered by existing Department of Transportation authorities). For some of the greatest national security harms arising from very large-scale computation, new actions and mechanisms will be required. For example, companies should be required to report the development or distribution of large AI computing clusters, training runs, and trained models (e.g., >1,000 AI chips, >10^26 operations, and >100 billion parameters, respectively). Government will need to step in if relevant computing clusters do not have adequate cybersecurity (such that adversaries could steal the resulting AI), or if the AI itself presents a national security threat (e.g., it can generate classified nuclear weapons information, design bioweapons or cyberweapons, or escape as a computer worm).

Threat assessments for these sorts of national security threats should be conducted at various points during AI development. These points should begin before training, because developing a dangerous model even for solely internal uses does not ensure perfect cybersecurity. The most-intensive analysis will have to begin after the model is trained, but updates and new discoveries should include a testing and evaluation component. A deployment license would be granted after the model passes training and pre-deployment threat assessments.

Kochhar, Andrew J. Lohn, and Osonde A. Osoba, *Air Dominance Through Machine Learning: A Preliminary Exploration of Artificial Intelligence–Assisted Mission Planning*, RAND Corporation, RR-4311-RC, 2020, https://www.rand.org/pubs/research_reports/RR4311.html; Lance Menthe, Li Ang Zhang, Edward Geist, Joshua Steier, Aaron Frank, Erik Van Hegewald, Gary Briggs, Keller Scholl, Yusuf Ashpari, and Anthony Jacques, *Understanding the Limits of Artificial Intelligence for Warfighters: Volume 1, Summary,* RAND Corporation RR-A1722-1, forthcoming.

While it is not possible to prevent bad actors from releasing deepfakes or similarly harmful media, it is possible to reduce the impact on Americans' information environment. The use of AI generation in political ads could require disclosure. There could be regulations restricting fine-tuning models to target a single individual. These requirements could be folded into licensing powerful AI models, or they could stand independently.

Only by making powerful AI trustworthy can it be deployed into high-stakes uses in the economy and national security. Therefore, the future of the AI industry hinges on whether powerful AI models are made safe and secure, which is why many firms are calling for government oversight to eliminate the temptation to compromise safety and security. By keeping safety and security requirements adaptable and allowing AI creators to come up with new ways to achieve safety, oversight will create innovative pressure to develop new methods for making AIs demonstrably safe and secure, which will create many positive spillovers and enable broader use of AI.

b. Would you support any sort of regulation to require disclosure of AI use in any work that involves a real individual and isn't simply a work of art?

Answer

AI is used to check spelling, validate grammar, draft documents, revise documents, read them out loud for editing, and more; AI use in work is too diffuse and too hard to detect to require disclosure in all instances. Where appropriate, requiring explicit notes in or adjacent to any photorealistic image that is AI generated would be useful for reducing the spread of AI-generated misinformation, as would requiring AI-generated audio to include a disclaimer in the clip. Such tags on whether a piece of content was produced by an AI or by an actual camera or microphone can be backed up by cryptographic signatures, which can confirm that a piece of media was made by specific real-world devices instead of a simulation.

Question from Rep. Jamaal Bowman

Question

AI is an inherently multidisciplinary field that requires a diverse set of expertise to develop fairly. In conversations with organizations that conduct interdisciplinary research, I've heard that it's notoriously difficult to obtain federal funding for their work. The National Science Foundation has previously offered grants for interdisciplinary research through programs like FairML. Do you think it would be valuable to expand public research funding for AI that encourages an interdisciplinary lens?

Answer

Private actors are investing heavily into creating new AI capabilities. In contrast, it will be vital to have interdisciplinary work that does not just ask how to do more but looks at what the societal impacts are likely to be, what as-yet unconsidered harms could arise, and what could be done about them. Expanding public research funding for AI safety; security; and ethical, legal, and societal implications will be critical to addressing society's challenges.

Question from Rep. Suzanne Bonamici

Question

An article in the Washington Post on June 19th detailed the disturbing world of AI-generated child sexual abuse images. Thousands of images circulate on dark web forums – images generated from known victims, or from benign images of children from photo-sharing sites and blogs.

a. No commitment to open-source transparency values should trump protecting children from such horrific exploitation. To what extent can developers of generative AI tools implement safeguards to prevent such despicable use of this technology? Is there any way to truly prevent these abuses once an open-source image generator has been released?

Answer

The default state is that a model that can produce both sexual images and images of children will be able to combine them. AI developers can implement safeguards by only releasing image models through an Application Programming Interface (API) or web interface and applying restrictions against misuse. However, if open-source image-generation models are released, neither of these solutions is viable.

b. The article explains that the proliferation of AI-generated images complicates law enforcement effort to identify victims in child sexual abuse images. How can developers assist in the effort to identify and eradicate these images and track down perpetrators?

Answer

Firstly, developers can watermark images. So long as models are closed-source, publication of methods to remove watermarks can be combated by modifying the watermark so that no new images can be produced with an easily removable watermark. Secondly, increasing the technical capacity of government investigators, through the methods discussed in my answer to Representative Lofgren's question, should help track offenders. Finally, automatic categorization and tagging of images could help identify child sexual abuse material across the internet.

Question from Rep. Emilia Sykes

Question

Should creators of open-source projects be regulated and held liable in the same manner as commercial services like ChatGPT, or should our laws draw a distinction between open-source software and commercial software?

Answer

Because safeguards can be removed from open-source models, AI models should undergo safety and security testing before being open-sourced. Commercial services that offer API access to underlying models can run tests to ensure that their models produce safe outputs in response to arbitrary inputs. Providers can track what outputs their models produce, pursuant to privacy agreements. Open-source software can always be modified by the end user, and there are no known methods to prevent end users from producing arbitrary outputs.