To: Stephanie Weiner, Acting Chief Counsel, NTIA
From: AI Impacts, Berkeley, CA
Re: AI Accountability Policy Request for Comment
Date: June 12, 2023

AI Impacts is a nonprofit research organization focused on the future of artificial intelligence and its associated risks and opportunities. In response to the Request for Comment, we have two general comments and some responses to specific questions.

I. Inadequate technical understanding

Because emerging artificial intelligence is poorly understood and our ability to predict how the technology will progress is limited, any accountability mechanisms that are created now will need to be flexible and kept up to date with progress in the field.

Holding parties accountable for their decisions requires a clear picture of which actions, outcomes, or policies are acceptable and which are not. However, the current state of the art in AI and AI safety is insufficient to provide clear boundaries on acceptable decision-making.[2][4] Examples of open problems or questions in artificial intelligence that bear on accountability include:

- Predicting which capabilities and which challenges to ensuring safety and fairness will emerge as existing systems are scaled up and new algorithms are developed.
- Determining in advance whether a particular training run for a machine learning model poses a serious risk to the public.
- Evaluating an AI system to determine whether it is safe to deploy, in a way that is robust to fine tuning, changes in operating context, and unexpected or malicious use.
- Predicting the effects that specific technical restrictions will have on future AI systems.

Mitigating risk through accountability will likely require risk assessment, but this cannot be done in a satisfactory manner without solving these or similar problems, and solving them may take considerable time, possibly years or decades. Given this, any accountability mechanisms created now should be built to accommodate future developments.

II. Accountability early in value chain

AI systems may pose risks that cannot be mitigated solely through post-training evaluation, suggesting that developers should be accountable for decisions made prior to training.

- Verification of the safety of a model may require that training be performed and monitored in a particular way.
- Evaluation itself may pose a risk. For example, GPT-4 (a state of the art large language model from OpenAI) deceived a human to solve a captcha during evaluation.[6]
- Training itself may pose a risk.

One reason to see the latter two as feasible is the generic tendency for agents to seek power.[11]

III. Responses to specific questions

Q2: The value of accountability measures will likely be their effects on internal processes or, given the fast pace of developments in AI, ensuring future adherence to safety requirements. Recent events such as Microsoft's Bing chat engaging in harmful speech after its release[7] and Meta's Large Language Model being leaked to the internet[3] suggest that AI developers' existing processes may be inadequate to ensure the safety of their products as AI capabilities continue to advance.

Q3: We do not see substantial barriers to furthering these goals simultaneously, and some measures will make progress on achieving most or all of them. For example, developing tools to verify that AI systems will reliably behave as intended and ensuring that these tools are implemented could improve accountability in high-stakes contexts, such as systems that will be given control of critical infrastructure, AI with influence over highly consequential decisions, or models with sufficient capability to pose a serious threat to the public.

Q9: The AI industry has no standard accountability mechanisms for the safe development and deployment of frontier AI models, but various researchers and organizations are doing preliminary work to develop and implement such mechanisms.

- DeepMind published a paper for evaluation of models, along with authors from several other AI labs and AI governance organizations, including OpenAI, the University of Oxford, and the Centre for the Governance of AI.[9]
- OpenAI is running a bug bounty program to identify security vulnerabilities with their systems.[5]
- The National Institute of Standards and Technology (NIST) released a risk management framework (RMF) for AI[10] and researchers at the Center for Long-Term Cybersecurity published supplementary guidance for the NIST RMF that addresses catastrophic risks.[1]
- ARC Evals (https://evals.alignment.org/) is researching methods of evaluating the capabilities and safety of AI models and has already done some work evaluating safety for OpenAI[6] and Anthropic.

Q31: Given the general lack of technical understanding for creating safe and fair AI (see Section I), building a strong accountability ecosystem may be well-served by government-funded research to better understand the technical hurdles we may encounter along the way. Specific research agendas that might be especially helpful for accountability include:

- Methods for evaluating AI systems and assessing risk. Robust methods can help regulators verify safety and help AI developers build trust with other stakeholers.
- Interpretability - Better tools for understanding how ML systems make decisions can improve our ability to evaluate them for safety and fairness.
- Hardware-level methods for verifying that a particular actor is using computing power as reported.[8]

Authors: Harlan Stewart and Dr. Richard Korzekwa

References

[1] Anthony M. Barrett et al. *Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks*. 2023. arXiv: 2206.08966 [cs.CY].

[2] Dan Hendrycks et al. "Unsolved Problems in ML Safety". In: *CoRR* abs/2109.13916 (2021). arXiv: 2109.13916. URL: https://arxiv.org/abs/2109.13916.

[3] Robert McMillan. "A Meta Platforms Leak Put Powerful AI in the Hands of Everyone". In: *The Wall Street Journal* (May 18, 2017). URL: https://www.wsj.com/articles/a-meta-platforms-leak-put-powerful-ai-in-the-hands-of-everyone-8b9f875a.

[4] Richard Ngo, Lawrence Chan, and Sören Mindermann. *The alignment problem from a deep learning perspective*. 2023. arXiv: 2209.00626 [cs.AI].

[5] OpenAI. *Announcing OpenAI's Bug Bounty program*. Apr. 2023. URL: https://openai.com/blog/bug-bounty-program.

[6] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[7] Kevin Roose. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled". In: *The New York Times* (Feb. 17, 2017). URL: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

[8] Yonadav Shavit. *What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring*. 2023. arXiv: 2303.11341 [cs.LG].

[9] Toby Shevlane et al. *Model evaluation for extreme risks*. 2023. arXiv: 2305.15324 [cs.AI].

[10] Elham Tabassi. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)". In: (2023).

[11] Alexander Matt Turner et al. *Optimal Policies Tend to Seek Power*. 2023. arXiv: 1912.01683 [cs.AI].