Comment Submitted to RFI Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

*Carson Ezell (cezell@college.harvard.edu) - Harvard University
*Stephen Casper (scasper@mit.edu) - MIT CSAIL

Charlotte Siegmann - MIT
Noam Kolt - University of Toronto
Taylor Lynn Curtis - MIT CSAIL
Benjamin Bucknall - Centre for the Governance of AI
Andreas Haupt - MIT
Kevin Wei - Harvard Law School
Jérémy Scheurer - Apollo Research
Marius Hobbhahn - Apollo Research
Lee Sharkey - Apollo Research
Satyapriya Krishna - Harvard University
Marvin Von Hagen - MIT
Silas Alberti - Stanford University
Alan Chan - Mila - Quebec AI Institute, Centre for the Governance of AI
Qinyi Sun - MIT
Michael Gerovitch - MIT
David Bau - Northeastern University
Max Tegmark - MIT
David Krueger - University of Cambridge
Dylan Hadfield-Menell - MIT CSAIL

Overview

Structured audits of AI systems are increasingly recognized as a way to increase accountability and identify risks from unsafe or societally harmful AI systems. Under Section 4.1(i)(C) of US Executive Order (EO) 14110 [34], NIST has been directed to develop "guidance and benchmarks for evaluating and auditing AI capabilities." We offer the following comment to the Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence [30] to accompany our recent paper, Black-Box Access is Insufficient for Rigorous AI Audits [14]. Our goal is to communicate the scientific consensus that (1) transparency regarding the access and methods used by auditors is needed to properly interpret audit results, and (2) white- and outside-the-box access allow for substantially more thorough assessments than black-box access alone.

Black-Box Access is Insufficient for Rigorous AI Audits

Recently, some developers of prominent state-of-the-art AI systems have kept most details of their models private [9]. To public knowledge, voluntary external audits of these systems have primarily involved analysis of the input/output behavior of models [3, 28, 35, 46]. This form of access in which auditors are only able to see outputs for given inputs is known as *black-box* access. Unfortunately, black-box access is very limiting for auditors. Some problems, such as anomalous failures, are difficult to detect with black-box access [24], and others, such as dataset biases, can be actively reinforced by testing data [42].

The ability to query a black-box system is useful, but many of today's evaluation techniques require access to weights, activations, gradients, or the ability to fine-tune the model [11]. *White-box* access refers to the unrestricted ability to observe a system's internal workings. It enables evaluators to apply more powerful attacks to automatically identify weaknesses [18, 37], study internal mechanisms responsible for undesirable model behaviors [21, 26], and identify harmful dormant capabilities through fine-tuning [39, 50]. Meanwhile, *outside-the-box* access involves additional contextual information about a system's development or deployment such as methodology, code, documentation, hyperparameters, data, deployment details, and findings from internal evaluations. It allows auditors to study risks that stem from methodology or data [6, 13, 29, 42] and makes it easier to design useful tests. This has led to a consensus in scientific discourse that white- and outside-the-box access allow for substantially more scrutiny than black-box access alone [1, 2, 10, 11, 41, 44].

Incorporating White- and Outside-the-Box Access into Practice

Absent proper guidance and regulatory action, black-box audits may become standard because they are precedented [32], existing calls for audits are often agnostic to form of access, and developers have incentives to limit external scrutiny. Industry actors have previously lobbied for limiting access given to auditors [19]. Here, we overview practical considerations involving white- and outside-the-box audits.

What kinds of systems should be considered for white- and outside-the-box audits? Prior work has argued that the rigor of AI audits should be proportional to the risks posed by the audited system [2, 40]. Examples of systems to consider for white- and outside-the-box audits may include:

- Models that qualify as dual-use foundation models under definition 3(k) in EO 14110 [34].
- Models that pose significant risks based on their application area—many high-risk applications were identified by the EU AI Act [16].
- Models that demonstrate high levels of agency [15, 43, 48, 49] which can be measured via performance on long-horizon tasks that require sophisticated planning [23, 45].

What skills and resources are needed for white- and outside-the-box audits? Some black-box AI evaluations can be conducted through very simple interfaces. However, some white-box (e.g., attacks, fine-tuning, interpretability) and outside-the-box (e.g., data or methodological analysis) techniques require extensive expertise and computing hardware.

What evaluation strategies do white- and outside-the-box access enable?

- White-box access allows for more powerful attack algorithms to design inputs which elicit harmful outputs from the system (e.g., instructions for committing crimes). This is largely due to how white-box access allows for gradient-based optimization [27].

- White-box access allows auditors to make stronger assurances against unforeseen failure modes by analyzing the system's robustness to perturbations to its internal state [25].
- White-box access allows auditors to fine-tune the system to assess risks from dormant capabilities and post-deployment modifications [39].
- White-box access allows auditors to search for explanations of behaviors and signs of undesirable internal mechanisms. For example, analyzing how models represent data involving different demographics could be used to assess its potential to discriminate [5].
- Outside-the-box access to training data can allow auditors to search for issues such as dataset biases [6], dataset poisoning [12], or copyright violations [22].
- Outside-the-box access allows auditors to analyze tradeoffs and risks taken by developers by assessing the methodology used to develop the system.
- Outside-the-box access to developers' internal evaluation results allows auditors to focus on a complementary set of evaluations.

How can leaks be avoided? The risk of leaks from auditors can be minimized through several technical, physical, and legal mechanisms. *Technical* solutions include providing auditors with de facto white-box access through application programming interfaces [7, 11, 36, 44]. *Physical* solutions can involve providing full white-box access through on-site secure research environments [20]. *Legal* solutions include formal training to protect confidentiality, non-disclosure clauses, clear terms of engagement in auditor-client contracts, and government standards, and have already been implemented in other industries with audits [4, 8, 17, 33, 38].

What disclosures are necessary to understand the limitations of an audit? Because the result of an audit can depend greatly on the methods that were used, the raw findings are insufficient to understand it alone. For regulators to properly interpret the outcome of an audit, they must also know what access was granted and what methods were used.

What kinds of public investments can help to develop tools and infrastructure for white- and outside-the-box AI audits? Auditors and developers alike benefit from improved techniques for evaluating, attacking, and interpreting AI systems. Public investments can help to facilitate further progress on these. First, government entities can offer support for scientific *research* into relevant techniques, such as the NSF's Safe Learning-Enabled Systems program [31]. Second, entities can develop secure evaluation *infrastructure*, such as the US National Deep Inference Facility [47], and subsidize usage costs for research with social benefits.

If there are any questions pertaining to our comment and/or recommendations, please contact Carson Ezell (cezell@college.harvard.edu) and Stephen Casper (scasper@mit.edu).

REFERENCES

[1] Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718* (2023).

[2] Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. 2023. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. (2023). arXiv:2311.14711 [cs.CY]

[3] Anthropic. 2023. Challenges in evaluating AI systems. (2023). https://www.anthropic.com/index/evaluating-ai-systems

[4] Compiled Auditing Standard ASA. 2006. Auditing standard ASA 210 terms of audit engagements.

[5] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48, 1 (2022), 207–219.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[7] Emma Bluemke, Tantum Collins, Ben Garfinkel, and Andrew Trask. 2023. Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases. (March 2023). https://arxiv.org/abs/2303.08956v2

[8] Public Company Accounting Oversight Board. [n. d.]. Driving Improvement in audit quality to protect investors. https://pcaobus.org/

[9] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. (Oct. 2023). http://arxiv.org/abs/2310.12941 arXiv:2310.12941 [cs].

[10] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (April 2020). https://doi.org/10.48550/arXiv.2004.07213 arXiv:2004.07213 [cs].

[11] Benjamin S Bucknall and Robert F Trager. 2023. Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements. (Oct. 2023). https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements/

[12] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2023. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149* (2023).

[13] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. (Sept. 2023). https://doi.org/10.48550/arXiv.2307.15217 arXiv:2307.15217 [cs].

[14] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv:2401.14446 [cs.CY]

[15] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. 651–666. https://doi.org/10.1145/3593013.3594033 arXiv:2302.10329 [cs].

[16] European Union. 2021. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[17] EY. 2019. EY Global Code of Conduct. Online. Retrieved from: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/generic/EY_Code_of_Conduct.pdf.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[19] Google. 2021. Consultation on the EU AI Act Proposal. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

[20] Susan Guthrie, Catherine A Lichten, Janna van Belle, Sarah Ball, Anna Knack, and Joanna Hofman. 2018. Understanding mental health in the research environment: A Rapid Evidence Assessment. *Rand health quarterly* 7 3 (2018), 2. https://api.semanticscholar.org/CorpusID:4610936

[21] Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786* (2023).

[22] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. *arXiv preprint arXiv:2310.13771* (2023).

[23] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. 2023. Evaluating Language-Model Agents on Realistic Autonomous Tasks. https://evals.alignment.org/language-model-pilot-report. (July 2023).

[24] Noam Kolt. 2023. Algorithmic black swans. *Washington University Law Review* 101 (2023).

[25] Nupur Kumari, Mayank Singh, Abhishek Sinha, Harshitha Machiraju, Balaji Krishnamurthy, and Vineeth N Balasubramanian. 2019. Harnessing the vulnerability of latent layers in adversarially trained models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2779–2785.

[26] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. *arXiv preprint arXiv:2401.01967* (2024).

[27] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. 2022. Adversarial attack and defense: A survey. *Electronics* 11, 8 (2022), 1283.

[28] METR. 2023. METR. https://evals.alignment.org/

[29] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[30] National Institute for Standards and Technology. 2023. Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11). https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the

[31] National Science Foundation. 2023. Safe Learning-Enabled Systems. https://www.nsf.gov/pubs/2023/nsf23562/nsf23562.htm

[32] Aaron L Nielson. 2018. Sticky Regulations. *U. Chi. L. Rev.* 85 (2018), 85.

[33] Electronic Code of Federal Regulations. 2023. Regulation M. Code of Federal Regulations. https://www.ecfr.gov/current/title-17/chapter-II/part-242/subject-group-ECFR3dd95cf4d3f6730 17 CFR Part 242.

[34] Office of the President of the United States. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[35] OpenAI. 2023. GPT-3.5 Turbo fine-tuning and API updates. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[36] Openmined. 2023. How to audit an AI model owned by someone else (part 1). *OpenMined Blog* (June 2023). https://blog.openmined.org/ai-audit-part-1/

[37] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, et al. 2016. Technical report on the cleverhans v2. 1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768* (2016).

[38] PCAOB. 2002. Sarbanes-Oxley Act of 2002. https://pcaobus.org/About/History/Documents/PDFs/Sarbanes_Oxley_Act_of_2002.pdf Public Law 107-204, 116 Stat. 745.

[39] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! *arXiv preprint arXiv:2310.03693* (2023).

[40] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

557–571.

[41] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153* (2023).

[42] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *Comput. Surveys* (2023).

[43] Yonadav Shavit, Cullen O'Keefe, Tyna Eloundou, Paul McMillan, Sandhini Agarwal, Miles Brundage, Steven Adler, Rosie Campbell, Teddy Lee, Pamela Mishkin, Alan Hickey, Katarina Slama, Lama Ahmad, Alex Beutel, Alexandre Passos, and David G Robinson. 2023. Practices for Governing Agentic AI Systems. *OpenAI* (Dec. 2023). https://openai.com/research/practices-for-governing-agentic-ai-systems

[44] Toby Shevlane. 2022. Structured access: an emerging paradigm for safe AI deployment. (2022). arXiv:2201.05159 [cs.AI]

[45] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324* (2023).

[46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023). arXiv:2307.09288 [cs.CL]

[47] United States National Science Foundation. 2023. National Deep Inference Facility for Very Large Language Models (NDIF). (2023).

[48] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).

[49] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

[50] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2023. Removing RLHF Protections in GPT-4 via Fine-Tuning. (2023). arXiv:2311.05553 [cs.CL]