**Trail of Bits**
85 Broad St, Fl 17
New York, NY 10004

## Trail of Bits's Response to NTIA AI Accountability RFC

***About Trail of Bits:*** *Since 2012, Trail of Bits has helped secure some of the world's most targeted organizations and devices. We combine high-end security research with a real-world attacker mentality to reduce risk and fortify code. We help our clientele—ranging from Meta to DARPA—lead their industries. Their dedicated security teams come to us for our foundational tools and deep expertise in reverse engineering, cryptography, virtualization, malware, and software exploits.*

***About the author:*** *Dr. Heidy Khlaaf is the Machine Learning (ML) Assurance Engineering Director at Trail of Bits and specializes in the evaluation, specification, and verification of complex or autonomous (e.g., ML) software implementations in mission-critical systems, ranging from UAVs to large nuclear power plants. Her expertise ranges from leading numerous system safety audits (e.g., IEC 61508, DO-178C) that contribute to the assurance of safety-critical software within regulatory frameworks and safety cases, to bolstering the dependability and robustness of complex software systems through techniques that identify and mitigate system and software risks. She is an author of the work cited in 75 by the NTIA RFC.*

Trail of Bits commends the National Telecommunications and Information Administration (NTIA) for fostering an open discussion on accountability and regulation through request for comments (RFC) on Artificial Intelligence ("AI") system accountability measures and policies. We offer recommendations informed by our expertise in cybersecurity and safety auditing of mission-critical software.

### Questions 1a, 1b, 1c, 1d, and 8 – On purpose and definitions of AI Accountability

We believe the definition of the term "audit," as defined in the supplementary information, is overly narrow and neglects important principles of auditing. We do not believe that the role of an audit is to assess performance against *accepted benchmarks* but rather verifiable claims. As discussed in the cited work 75, the scope carried out should be relevant to a regulatory, safety, ethical, or technical claim, for which stakeholders may be held accountable. Accepted benchmarks may be scoped too narrowly or may not have any relevance to the claim being made. That is, benchmarks should not inform AI accountability measures but rather should be used as evidence or a mechanism toward the verifiability of a specific claim. Claim-oriented approaches have been developed in order to

structure arguments about the safety of engineered systems in safety-critical domains[1], as discussed further in our responses to Questions 9, 20, and 30.

We note that the majority of algorithmic assessments introduced for AI/ML-based systems aim to audit general properties of a system without considering its claims, and thus a lack of an operational envelope. The lack of a defined operational envelope for the deployment of generative models has rendered the evaluation of their risk and safety intractable due to the sheer number of applications and, therefore, the potential risks posed. In [2], we propose the integration of Operational Design Domains (ODDs) as first introduced by the National High Traffic Safety Administration (NHTSA)[3] into a risk framework, where we define a novel ODD taxonomy relevant to the use of AI technologies, including generative models. The purpose of an ODD is to describe the specific operating conditions under which an AI-system is designed to properly behave, thus outlining the safety envelope against which system hazards and harms can be determined. The use of ODDs can guide in understanding the constraints under which the AI system may no longer behave as intended or how it can escape its designated safety envelope.

Finally, the definition of the term "audit" forgoes independence, an attribute crucial to established auditing practices in other fields. Take, for example, the following clause from IEC 61508:2010, a functional safety standard applicable to safety-critical domains:

*IEC 61508-4:2010, 3.8.4 - systematic and independent examination to determine whether the procedures specific to the functional safety requirements to comply with the planned arrangements are implemented effectively and are suitable to achieve the specified objectives. Note: A functional safety audit may be carried out as part of a functional safety assessment.*

Independence allows the public to trust in the accuracy of the results and the integrity of the resulting outcomes. We encourage the NTIA to consider independence as a pillar of trustworthiness and auditing when considering system accountability.

**Question 1e – Can AI accountability practices have meaningful impact in the absence of legal standards and enforceable risk thresholds? What is the role for courts, legislatures, and rulemaking bodies?**

Only insofar as assessing already specified claims, to which developers can be held accountable, and providing argumentation that addresses the evaluation and risk assessment of the system and the role of the different subsystems in achieving trustworthiness.

---

[1] Bloomfield, R., Khlaaf, H., Ryan Conmy, P., and Fletcher, G., "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy". Computer, vol. 52, no. 9, pp. 82-89, Sept. 2019, doi: 10.1109/MC.2019.2914775.

[2] Khlaaf, H., "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems". Trail of Bits, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

[3] "A Framework for Automated Driving System Testable Cases and Scenarios". National Highway Traffic Safety Administration. DOT HS 812 623. https://rosap.ntl.bts.gov/view/dot/38824.

When establishing how much the technologies can and need to be trusted, there must be an articulated claim of what they will be used for and the risk threshold accepted by civil society that must be determined in part by legislatures, rulemaking bodies, and regulators. If the vision of how something will be used is not clearly formulated, we cannot use accountability practices to determine how much we need to trust it or what the risks are. Furthermore, the absence of legal standards and enforceable risk thresholds only entails an implicit acceptance of thresholds as determined by those with self-interest in the development and deployment of AI-based systems.

**Question 2 – Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?**

We believe accountability mechanisms are intended both to promote trust among external stakeholders and to change internal processes to allow AI-based systems to be amenable to audits and assessments. Certifications, audits, and assessments can ensure trustworthiness that would promote the public's confidence that the entities building AI—including civil society, governments, the private sector, and other stakeholders—are doing so responsibly, with positive intentions, competence, and accountability. Internal processes can be modified to allow for audit trails that can improve the verifiability of claims about engineered systems. This is further discussed in the cited work 75 in "3.1 Audit Trails."

**Question 5 – Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?**

We believe such accountability claims can be made only with due consideration of the scope and the operational envelope in which a general-purpose AI model is intended to function, as noted by our response to Questions 1 and 8, and outlined in [4].

**Question 7 – Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?**

The impact of accountability mechanisms on AI innovation can only be considered relative to a level of risk that is deemed acceptable. That is, to consider accountability mechanisms a hindrance, it must be possible to demonstrate that the cost involved in reducing risks would be grossly disproportionate to the benefit gained. This concept is in fact known as "As Low As Reasonably Practicable" (ALARP) within the safety-critical domain[5]. No risk thresholds for AI have been defined by legislation or regulation, nor has evidence been provided regarding the cost of implementing

[4] Khlaaf, H., "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems". Trail of Bits, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

[5] "Tolerability of Risk the ALARP Philosophy". IAEA NGKB Nuclear Graphite Knowledge Base. Retrieved June 7, 2023, from https://nucleus.iaea.org/sites/graphiteknowledgebase/wiki/Guide_to_Graphite/Tolerability%20of%20Risk%20the%20ALARP%20Philosophy.aspx

accountability mechanisms by those developing AI-based systems to be able to make such a determination.

Furthermore, it is difficult to ascertain the technical barriers that can impede the implementation of accountability mechanisms given that existing AI-based systems do not follow rudimentary software safety and security best practices (e.g., IEC 61508, NIST 800-154; see Questions 9 and 20 below). Leveraging fundamental best practices[6] as a first step can serve accountability goals at the same cost as following software best practices, which can then enable the development of AI-specific accountability mechanisms.

### Questions 9, 20, and 30 – On existing accountability mechanisms, their use, and role

In safety-critical and defense domains, claim-oriented or goal-based approaches are used in order to structure arguments about the safety of engineered systems, including autonomous and AI-based systems[7]. These approaches are predominantly known as safety or assurance cases. A safety case is a documented body of evidence that provides a convincing and valid argument regarding a top-level claim (such as the safety of an autonomous vehicle as defined in UL 4600[8]), and presents a structured justification in support of that claim to decide the status of it. Safety cases are often required as part of a regulatory process. For example, the FDA requires infusion pump manufacturers to submit safety cases as part of the 510(k)s.

A certificate of safety or a license is then granted only when a regulator is satisfied by the argument presented in a safety case. The goal-based approach and fluidity of a safety case allows licensees to determine the assurance activities that must be carried out in accordance with a regulator's safety goals or principles. Licensees are then responsible for ensuring that their use of a technology complies with these principles by conducting or commissioning assessments of their systems. This process leads to a documented formal qualification of a system for its intended application, backed by evidence, that can be presented to a regulator.

When determining the safety justification of software-based systems within a safety case, it is typically split across two stages: production excellence (i.e., accountability-by-design), in which the quality of the design and development processes is assessed, and independent assessment, which requires a thorough, independent examination of the device and/or its software. Production excellence is typically assisted by evidence of the systematic application of national and international standards (i.e., prescriptive approaches). IEC 61508, UL 4600, IEC 61513, and DO-178C are typical of the standards recommended for this role. We refer to these standards for a detailed review of

---

[6] Kroll, J. A., "Outlining traceability: A principle for operationalizing accountability in computing systems". Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 758–771. https://doi.org/10.1145/3442188.3445937

[7] Bloomfield, R., Khlaaf, H., Ryan Conmy, P., and Fletcher, G., "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy". Computer, vol. 52, no. 9, pp. 82-89, Sept. 2019, doi: 10.1109/MC.2019.2914775.

[8] "UL 4600: Standard for Evaluation of Autonomous Products, Edition 3". Underwriters Laboratories, March 2023.

records and other documentation required for systems to support AI accountability, as we believe that AI systems should be categorized as an extension of software-based systems given the identical mechanisms of their development. Current AI-based systems do not possess any unique software components that warrant a generalized licensing scheme that would not heavily impede the use of software as a whole. Indeed, any implementation of such a scheme would likely result in significant overreach due to the broad definition and software components of AI systems. A further literature review on accountability mechanisms, including regulation and assessments, required for software-based systems across numerous safety-critical domains can be found in [9].

We believe the above processes would be too rigorous for non-mission-critical AI applications, and that AI regulatory policies should generally mirror the practices of existing sectors in which they are deployed. This includes inputs to audits or assessments and mandating accountability measures, including compliance with existing regulatory standards and practices throughout a system's lifecycle to provide assurance of the final design. Overall, AI accountability policies and regulations should largely be sectoral, and further regulation should be defined for novel domain areas where AI may produce novel harms (e.g., bias and discrimination in facial recognition). As noted in our answer to Question 1, we believe that by defining a more concrete operational envelope (e.g., through a sector-specific and AI-based ODD[10]), developers and regulators can better assess potential risks and required safety mitigations for AI-based systems.

**Question 10 – What are the best definitions of terms frequently used in accountability policies, such as fair, safe, effective, transparent, and trustworthy? Where can terms have the same meanings across sectors and jurisdictions? Where do terms necessarily have different meanings depending on the jurisdiction, sector, or use case?**

Although there can be well-established overarching definitions—for example, those provided in OECD's AI Principles, or in the European Union's Ethics Guidelines for Trustworthy AI—we believe these terminologies can be fully realized and defined only when they are given an application and scope. As we discussed in our answers to Questions 1 and 8, the use of ODDs can guide in understanding the constraints under which the AI system may no longer behave as intended or how it can escape its designated safety envelope.

**Question 15c – How should vendors work with customers to perform AI audits and/or assessments? What is the role of audits or assessments in the commercial and/or public procurement process? Are there specific practices that would facilitate credible audits (e.g., liability waivers)?**

We believe this question is best answered by the work in [11].

---

[9] Butler, E., Fletcher, G., George, S., Guerra, S., and Khlaaf H., "Cots Digital Devices In Safety Critical Industries – Use and Licensing". Energiforsk AB, November 2019, ISBN 978-91-7673-627-2.

[10] Khlaaf, H., "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems". Trail of Bits, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

[11] Clark, J., and Gillian K. H., "Regulatory markets for AI safety." arXiv preprint arXiv:2001.00078, 2019.

## Questions 16 a, b and 17 - On AI Accountability focus and scope

We believe the posed dichotomy of technical versus socio-technical assessment fails, historically and practically, to recognize a nuance in the consideration of distinct assessment approaches. Indeed, it is the case that the "most useful audits and assessments of these systems … should extend beyond the technical to broader questions about governance and purpose"; however, technical assessments are not intended to support purely technical goals, but rather claims regarding the holistic behavior of the system and how a system may technically achieve them. That is, technical assessments are also a necessary tool in supporting socio-technical, legal, or regulatory claims regarding the fitness of the system. As noted in our answers to Questions 9, 20, and 30, the mixture of claim-oriented approaches and prescriptive standards in safety-critical domains provides such an example. The use of varying techniques (both technical and socio-technical) across a system's lifecycle is not competing but rather serves different purposes[12] in substantiating system claims, requiring that both be used within an assessment scope.

Accountability mechanisms should be communicated through the context (e.g., operational envelope or ODD) that defines the scope of the AI claims. A claim may hold true only within the boundaries of that scope, which must be specified. Regarding scoping of risks, it is pertinent to assess model capabilities through application-specific evaluation benchmarks to inform risk assessments. It may seem counterintuitive to carry out a capabilities evaluation before a risk assessment. However, traditional risk assessments require implicit assumptions and knowledge regarding a prospective system's capacities, limitations, and failure modes (which in turn inform possible harms a system may pose) [13]. In the case of Large Language Models (LLMs), for example, and more generally, generative AI, these capabilities and failure modes are not yet fully understood. The acceptance or mitigation of identified hazards and harms within a risk assessment must be evaluated based on performance criteria that define the tolerable risk allowed. Although works in [14] and [15] define novel AI-specific hazard categories and a baseline operational envelope or context under which harms must be considered, risk thresholds must be determined in part by legislatures, rulemaking bodies, and regulators (as noted in our answer to Question 1e). In low-harm applications or domains that require no further regulation, this may be determined by business operations.

---

[12] Bloomfield, R., Khlaaf, H., Ryan Conmy, P., and Fletcher, G., "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy". Computer, vol. 52, no. 9, pp. 82-89, Sept. 2019, doi: 10.1109/MC.2019.2914775.

[13] Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., & Brundage, M., "A hazard analysis framework for code synthesis large language models". arXiv. http://arxiv.org/abs/2207.14157

[14] Khlaaf, H., "Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems". Trail of Bits, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

[15] Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., & Brundage, M., "A hazard analysis framework for code synthesis large language models". arXiv. http://arxiv.org/abs/2207.14157

## Bibliography


- “A Framework for Automated Driving System Testable Cases and Scenarios”. National Highway Traffic Safety Administration. DOT HS 812 623. https://rosap.ntl.bts.gov/view/dot/38824.
- Bloomfield, R., Khlaaf, H., Ryan Conmy, P., and Fletcher, G., "Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy". Computer, vol. 52, no. 9, pp. 82-89, Sept. 2019, doi: 10.1109/MC.2019.2914775.
- Butler, E., Fletcher, G., George, S., Guerra, S., and Khlaaf H., “Cots Digital Devices In Safety Critical Industries – Use and Licensing”. Energiforsk AB, November 2019, ISBN 978-91-7673-627-2.
- Clark, J., and Gillian K. H., "Regulatory markets for AI safety." arXiv preprint arXiv:2001.00078, 2019.
- Khlaaf, H., “Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems”. Trail of Bits, 2023. https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.
- Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., & Brundage, M., “A hazard analysis framework for code synthesis large language models”. arXiv. http://arxiv.org/abs/2207.14157
- Kroll, J. A., “Outlining traceability: A principle for operationalizing accountability in computing systems”. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 758–771. https://doi.org/10.1145/3442188.3445937
- “Tolerability of Risk the ALARP Philosophy”. IAEA NGKB Nuclear Graphite Knowledge Base. Retrieved June 7, 2023, from https://nucleus.iaea.org/sites/graphiteknowledgebase/wiki/Guide_to_Graphite/Tolerability%20of%20Risk%20the%20ALARP%20Philosophy.aspx
- “UL 4600: Standard for Evaluation of Autonomous Products, Edition 3”. Underwriters Laboratories, March 2023.