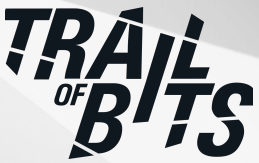

A Comprehensive Risk Assessment Framework for AI Assurance in Ethical, Legal, and Societal Domains

Prepared in Response to:

Ethical, Legal, and Societal Implications (ELSI) of Emerging Technologies RFI

DARPA-SN-23-68@darpa.mil

June 9, 2023

Prepared By:

Dr. Heidy Khlaaf | *Engineering Director, ML Assurance Practice*

heidy@trailofbits.com

Michael D. Brown | *Principal Security Engineer, Research Practice*

michael.brown@trailofbits.com

Table of Contents

1. Introduction

Emerging Artificial Intelligence and Machine Learning (AI/ML) technologies such as generative models (i.e., Large Language Models or LLMs) have recently demonstrated a significant increase in capability that has captured the national interest. Deploying these systems without clearly defined requirements and risk analyses has led to safety hazards and novel ethical, legal, and socio-economic (ELS) harms [4, 8, 12]. Few, if any, methodologies to systematically identify and address these harms have been developed, and today's AI risk frameworks and metrics are poorly suited for quantifying novel AI failure modes and harms. Inconsistent use of baseline terminology has led to contradictory approaches that conflate requirements engineering with safety measures, resulting in drastically different outcomes when constructing risk and hazard assessments.

Effective development, transition, and deployment of disruptive AI/ML technologies, particularly for the US Government, require a comprehensive and robust approach to AI assurance that proactively identifies and mitigates safety hazards and ELS harms. We propose a novel AI assurance and risk assessment methodology based on Operational Design Domains (ODDs), a concept initially introduced for automated driving systems. Our work overcomes the limitations of applying existing approaches [9, 13, 14, 18] adopted from hardware, software, security, and systems safety communities to AI systems and aims to help developers and auditors build confidence that AI/ML systems have fully addressed their safety and ELS risks.

We address RFI topic areas 1, 2, 5, and 8 in our response, organized as follows. Section 2 discusses how inconsistent use of baseline terminology undermines AI assurance and risk assessment. Section 3 reviews the limitations of applying hardware, cybersecurity, system safety risk, and software safety techniques to AI assurance and risk assessment. Finally, Section 4 proposes a novel assurance approach that outlines a comprehensive risk assessment framework that overcomes many of the discussed limitations.

2. Distinguishing Value Alignment and Safety

Today's methods for assessing AI risk come from well-established system safety techniques, which inconsistently use and misconstrue the meaning of several important terms. Some methodologies define "safety" as the prevention of failures due to accidents [3, 15], while others measure safety in terms of the AI system's alignment with human-oriented values and goals (e.g., fairness, harmlessness) [11, 6]. To their detriment, value alignment measures are subjective and conflate *safety properties* with *system requirements*. Compare the following established definitions:

- **Value Alignment** [6]: AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

- **System Requirement**: A statement that translates or expresses functionality to satisfy intent or stakeholders' needs.

Given that *intent* and *stakeholders' needs* are subjective human values, the term "Value Alignment" is a specific type of system requirement. This confusion of terminology is critical, considering the definition of "safety" often deployed for safety-critical systems:

- **Safety**: To prevent a system from impacting its environment in an undesirable or harmful way, typically aiming to protect human lives, natural environment, or monetary assets.

Safety concerns are derived from harms posed by a system meeting *its specification*. Conflating requirements engineering with safety measures creates a false equivalence between *system safety* and *a system meeting its intent*. This stunts assurance and risk assessment and allows identifiable hazards and harms to go unmitigated. Safety must center on the lack of harm that may arise due to the intent itself or failures arising in an attempt to meet said intent (e.g., implementation failures). Thus, AI assurance and risk methodologies that root safety in alignment do not adequately address exploring the ELS harms [4, 17] that will arise from an AI system meeting well-intentioned specifications (as they do with safety-critical systems), especially given their complexity, scale, nondeterminism, and unknown failure modes.

3. Pitfalls in Existing Adoptions and Approaches

Prior attempts to apply well-established safety and security techniques AI/ML systems have serious limitations. For example, risk assessment techniques [14] adopted from hardware safety (e.g., FMEA) are unsuitable for AI-based systems as they measure safety properties of systems under the assumption of random failures. This is not conducive to uncovering the design issues that directly lead to systematic failures in AI/ML systems.

Additionally, works such as [9, 13, 18] have adopted cybersecurity practices (e.g., threat modeling, red teaming) to uncover hazards or harms for AI-based systems. These approaches aim to prevent external adversarial agents from impacting a system in a harmful way. In contrast, safety aims to *prevent a system from impacting its environment* in a harmful way (e.g., protect human lives, the environment, etc). Therefore, safety risk frameworks may be more appropriate for exploring AI/ML systems' harms than threat modeling, which aims to protect a system from its external environment.

We argue for building AI-specific risk frameworks on more relevant system-level risk assessment frameworks [19], such as MIL-STD-882e [16]. Such general frameworks have an expanded scope to address hazards, harms, and systematic considerations crucial to AI/ML systems, such as general system failures and emergent behaviors. Still, there are limitations to these approaches that must be addressed, namely (1) the stochastic (and

non-deterministic) nature of AI/ML models makes it difficult to determine tolerable risk for harms, and (2) techniques to assess AI/ML system robustness equivalent to those used in other domains (e.g., static analysis and formal verification of software) are still under development [7].

4. Unifying Risk Assessment and Safety Justification

We propose a novel systematic AI assurance and risk assessment methodology that adapts system safety engineering approaches and addresses the core problem of operationalizing risk modeling to construct comprehensive AI assurance claims. We employ a system-level risk assessment to define criteria for the tolerable risk allowed and guide system design to reduce the frequency of identified harms, helping developers and auditors build confidence that an AI-based system has addressed its safety and ELS risks throughout implementation and deployment [2].

Our methodology integrates Operational Design Domains [1], initially introduced for automatic driving systems (ADS), into risk assessments for general AI-based systems. ODDs describe the operating conditions for which an AI system is designed to behave properly, outlining the safety envelope for which system hazards and harms can be determined against. Defining a concrete operational envelope helps developers and auditors assess potential risks and required safety mitigations for AI/ML systems by defining the constraints under which the system no longer behaves as intended or can escape its designated safety envelope [19]. For example, an ADS deemed safe for highway driving in clear weather may not be safe on city roads or in adverse weather conditions.

To integrate ODDs into a risk framework, it is necessary to define a novel ODD taxonomy relevant to AI technologies, including general multi-modal models. We define a baseline taxonomy [19] subdivided into categories including Application/Domain, Users/Agents, Vector, Protected Characteristics, and Assets as is done in [1]. Our ODD taxonomy is further divided into subcategories that we omit for brevity. A full listing and diagram of our novel AI ODD taxonomy can be found in Appendix A, Figure 2.

With this taxonomy, we will demonstrate how an ODD can be used to systematically identify ELS risks using a risk framework defined in [10], relevant resources of which are provided in Appendix B. ODDs can be seen as a consideration of various permutations of the categories to effectively explore a wide range of scenarios and their associated risks. With an identified application domain and a system abstraction scope defined for the risk assessment, assessors should enumerate through all categories and subcategories and consider all hazards that may arise due to their interaction. This can be best illustrated with the example figure below, where each hazard has an associated field for each ODD category. We provide three distinct examples of hazard entries across varying application domains and system abstraction levels.

Hazard Source	Hazard Description	Trigger Event	Application/Domain	Users/Agents	Vector	Protected Characteristic	Assets	Potential Effects
System Design - Model Training	Subpopulations not appropriately identified or represented within distributions of the dataset.	Use of CNN skin cancer classifier on skin on a non-white skintone.	Human Health	End user	Data - Training	Race	Data	Increased risk of false negative, or misclassification of skin cancer as a benign or not present.
System Implementation	Not a Number (NaN) value propagates to speed limit value in autonomous vehicle planning module.	Lack of float precision and calculation leading to Not a Number (NaN).	Automotive & Transportation	Model user	Model - Source Code	N/A	AI system components and subcomponents	NaN causing uncontrolled acceleration leading to increased risk of physical harm, injury, or death.
Human Rights	Exposure to toxic datasets or model inputs that promote hate crimes towards a labeler's gender identity.	Labeling and fine-tuning datasets for violence, hate speech, etc., to train model to learn to detect those forms of toxicity.	Marketing, Advertising, and Microtargeting	HITL - Fine Tuning & Corrective	Data - Training	Gender identity	Human Resources	Mental distress, grief, and potential impact to long-term mental illnesses (e.g., PTSD).

Figure 1: Demonstration of hazard entries within a hazard analysis using ODD categories (colored) as fields for consideration.

Note that in Figure 1 above, "Hazard Risk Index (HRI)" and "Mitigation" fields have been omitted for presentation purposes. The HRI (see Appendix B, Table 3) for each hazard assists in understanding how the risks compare to each other, and the priority with which each hazard must be controlled. "Mitigations" describe specific actions that would reduce the associated HRI for a hazard. Action points arising from mitigations comprise the associated system safety requirements accompanying system or operational requirements. With each set of mitigations implemented, the HRIs should be recalculated, and the process should be repeated until all undesirable risk is eliminated for the safe deployment of an AI-based system. A risk template for use with the corresponding ODD categories is available in Appendix B, Table 3.

5. Conclusion

Building and deploying safe AI/ML systems that fully mitigate ELS harms requires an operationally-focused, comprehensive, and robust assurance and risk assessment approach. Unfortunately, misalignment of key terminology (e.g., "safety" and "alignment") and fundamental limitations of existing approaches used in the hardware, software, security, and systems safety engineering communities prevent the direct application of prior work to AI/ML systems. To overcome these problems, we propose a novel assurance and risk assessment approach that integrates ODDs via a novel taxonomy relevant to AI technologies, including general multi-modal models. The taxonomy is subdivided into categories that allow for exploring a wide range of scenarios and their associated risks. By defining a more concrete operational envelope, we believe developers and auditors can better assess potential risks and required safety mitigations for AI-based systems. We invite the reader to consult the expanded version of this work [19] for more detail and resources.

Bibliography

1. 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
2. ALARP - As low as reasonably practicable. Health and Safety Executive (UK Gov). https://www.hse.gov.uk/comah/alarp.htm
3. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. http://arxiv.org/abs/1606.06565
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922
5. Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research 81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html
6. Brown, D. S., Schneider, J., Dragan, A. D., & Niekum, S. (2021). *Value alignment verification*. arXiv. http://arxiv.org/abs/2012.01557
7. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., ... Anderljung, M. (2020). *Toward trustworthy AI development: Mechanisms for supporting verifiable claims*. arXiv. http://arxiv.org/abs/2004.07213
8. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230
9. Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*. arXiv. http://arxiv.org/abs/2209.07858
10. Khlaaf, H., Mishkin, P., Achiam, J., Krueger, G., & Brundage, M. (2022). *A hazard analysis framework for code synthesis of large language models*. arXiv. http://arxiv.org/abs/2207.14157
11. Langosco, Lauro Langosco Di; Koch, Jack; Sharkey, Lee D; Pfau, Jacob; Krueger, David (July 17, 2022). "Goal misgeneralization in deep reinforcement learning." International Conference on Machine Learning. Vol. 162. PMLR. pp. 12004–12019.
12. Mittelstadt, Brent and Wachter, Sandra and Russell, Chris (2023), The Unfairness of Fair Machine Learning: Leveling down and strict egalitarianism by default. https://ssrn.com/abstract=4331652

13. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red teaming language models with language models*. arXiv. http://arxiv.org/abs/2202.03286
14. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873
15. Raji, I., & Dobbe, R. (2020). *Concrete problems in AI safety, revisited*. ICLR workshop on ML in the real world.
16. Mil-std-882e, Department of Defense standard practice system safety (2012). US Department of Defense (2012).
17. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. arXiv. http://arxiv.org/abs/2112.04359
18. Yee, Kyra, and Irene F. Peradejordi. Sharing learnings from the first algorithmic bias bounty challenge. Twitter Blog. https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge.
19. Khlaaf, Heidy. Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems, Trail of Bits, 2023.

Appendix A - Novel AI System ODD Taxonomy

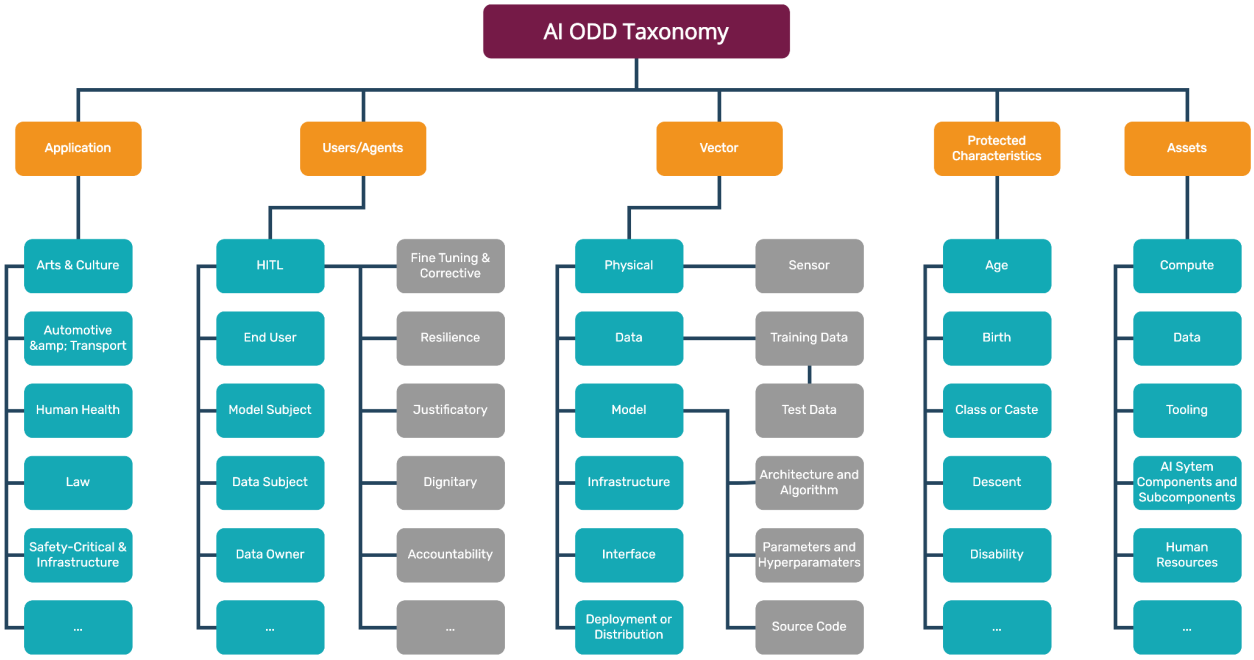

Figure 2: AI ODD Taxonomy with the baseline categories and sample subcategories for illustrative purposes

Appendix B - Framework Resources

Description	Category	Definition (Mapped to Table 2)
Catastrophic	1	Death, permanent total disability, direct harm, significant system or asset loss, or irreversible significant environmental impact.
Critical	2	Incitement, manipulation, radicalization, or discrimination that may result in mental or physical harm. Cause of consequential error to many individuals, reversible significant environmental impact, or considerate asset loss.
Major	3	Injury or cause of consequential error to a few individuals, reversible environmental impact, or moderate asset loss.
Minor	4	Injury or cause of consequential error not resulting in any long term harm, minimal environmental impact, or negligible asset loss.

Table 1: Hazard Severity Categories associated with the use of general multi-modal models [10, 16]

Hazard Frequency	Catastrophic	Critical	Marginal	Negligible
(A) Frequent	1A	2A	3A	4A
(B) Probable	1B	2B	3B	4B
(C) Occasional	1C	2C	3C	4C
(D) Remote	1D	2D	3D	4D
(E) Improbable	1E	2E	3E	4E

Table 2: Hazard Risk Index [16] considering hazard frequency against its severity category

ID	Hazard Source	Hazard Description	Trigger Event	Application/Domain	Users/Agents	Vector	Protected Characteristics	Assets	Potential Effects	HRI	Recommend Mitigations	HRI (post)
H1	System Design - Model Training	Subpopulations not appropriately identified or represented within distributions of the dataset.	Use of CNN skin cancer classifier on skin on a non-white skin tone.	Human Health	End user	Data - Training	Race	Data	Increased risk of false negative, or misclassification of skin cancer as a benign or not present.	2B		

Table 3: Risk assessment template with ODD categories as fields

Appendix C - About Trail of Bits

Founded in 2012 and headquartered in New York, Trail of Bits provides technical security assessment, research, and engineering services to some of the world's most targeted organizations. We combine high-end security research with a real-world attacker mentality to reduce risk and fortify software and AI/ML systems. With 100+ employees around the globe, we've helped secure critical software elements that support billions of end users (e.g., Kubernetes, Linux kernel, AlgoVPN) and has developed novel, open-source security tools under several DARPA programs (CHESS, AMP, SafeDocs, SIEVE, V-Spells, PACE).

We specialize in software testing, code review, and threat modeling projects supporting client organizations in the technology, defense, and finance industries, as well as government entities. Notable clients include HashiCorp, Google, Microsoft, Western Digital, and Zoom. Trail of Bits' machine learning (ML) assurance practice creates tools and techniques for exploring attack surfaces and failures in ML-based systems that can lead to the degradation of model performance, exploitation of assets, or manipulation of outputs.

Trail of Bits' research and engineering practice specializes in the research and development of novel software analysis and transformation capabilities, including those that are based on and target AI/ML models. Notable clients include DARPA, the Office of Naval Research, and the US Army RDECOM. To date, Trail of Bits has created and maintains more than 200 free and open-source tools (available in our GitHub repositories) and offers our research and engineering services for the public and private sectors.

In recent years, Trail of Bits consultants have showcased cutting-edge research through presentations at IEEE S+P, AISec, CanSecWest, HCSS, Devcon, LangSec, the Linux Security Summit, the O'Reilly Security Conference, PyCon, RWC, REcon, and SummerCon.

We maintain an exhaustive list of publications at https://github.com/trailofbits/publications, with links to papers, presentations, public audit reports, and podcast appearances.

To keep up to date with our latest news and announcements, please follow @trailofbits on Twitter and explore our public repositories at https://github.com/trailofbits. To engage us directly, visit our "Contact" page at https://www.trailofbits.com/contact, or email us at info@trailofbits.com.

Trail of Bits, Inc.
228 Park Ave S #80688
New York, NY 10003
https://www.trailofbits.com
info@trailofbits.com