Envisioning a Healthy and Thriving Ecosystem for Assessment of Foundation Models

Ross Gruetzemacher[1,2], **Iyngkarran Kumar** [2], **Kyle Kilian** [2]

[1]Wichita State University
[2]Transformative Futures Institute
ross.gruetzemacher@wichita.edu, iyngkarrankumar@transformative.org, kyle@transformative.org,

Abstract

In this extended abstract we describe three issues hindering the development of a healthy and thriving ecosystem for assessing foundation models. We then propose actionable steps to address these issues and move toward a healthy and thriving ecosystem. We do not suggest these steps to be a solution, rather, we hope that this can be the catalyst for discussion about envisioning the long-term health and flourishing of the AI assessment ecosystem as assessment of capabilities becomes increasingly important in regulating advanced AI systems in order to mitigate societal harms.

Introduction

Foundation models are general-purpose AI systems trained on broad data that are easily adapted to a wide range of downstream tasks with minimal domain-specific training (Bommasani et al. 2021). Increasingly large models have demonstrated ever more powerful general-purpose abilities (Bubeck et al. 2023; Arcas and Norvig 2023). These characteristics are expected to drive economic prosperity and growth, advance science, and transform society (Gruetzemacher and Whittlestone 2022); however, foundation models also create new challenges to ensuring that they are used safely and responsibly. Consequently, mitigating the societal harms of large-scale frontier AI deployment will require robust and coordinated testing, evaluation, verification, and validation (TEVV; (NIST 2023)).

Some specific capabilities of foundation models are particularly concerning and will require novel TEVV approaches. For example, autonomous replicating and adaptation (ARA) could have difficult to anticipate consequences (Kinniment et al. 2023), and minimal fine-tuning models already fine-tuned by developers to mitigate undesirable behavior often leads to models reverting to their initial states, effectively unlearning the desirable behaviors (Jain et al. 2023), even unintentionally (Qi et al. 2023).

TEVV is a broad topic, and managing foundation models' risks and harms involves unique challenges assessing models' potentially harmful capabilities. Thus, we focus on capabilities assessment—the T and E of TEVV. Assessment of models' capabilities is challenging and requires novel approaches, so it is critical to identify steps to foster a healthy, thriving ecosystem for assessing foundation models' capabilities to ensure efforts to mitigate harms are most effective.

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

Problems with the Present Ecosystem

There are three major problems with the current assessment ecosystem. We briefly describe each below.

First, there are several competing perspectives on the assessment of foundation models' capabilities. These different schools of thought have emerged from different fields and are inherently divergent, pushing the current ecosystem toward fragmentation. We see four unique schools: the TEVV school, the aggregate benchmark school, the 'evals' school, and the cognitive school. The TEVV school is exemplified by NIST, and the work it has done on certifying autonomous systems (NIST 2023), as well as other software systems (NIST 2024). The aggregate benchmark school is exemplified by benchmarks such as SuperGLUE (Wang et al. 2019), MMLU (Hendrycks et al. 2020), and HELM (Liang et al. 2022). The 'evals' school focuses on extreme capabilities evaluations (Shevlane et al. 2023), such as dangerous emergent capabilities (Wei et al. 2022), often using adversarial testing or red teaming; this school is commonly associated with Apollo Research (Sharkey et al. 2024), METR (Kinniment et al. 2023), or AI developers like Anthropic or Google (Weidinger et al. 2023). Last, the cognitive school derives from psychology (e.g., psychometrics) and is associated with efforts to identify models' capabilities and risks more fundamentally (Zhou et al. 2023; Dentella et al. 2023). Methods from this school are sometimes considered 'evals', too. Then there are groups that do not neatly fit into just one of the schools, such as Scale, which proposed its own *Test and Evaluation Vision* (ScaleAI 2023).

Each of these schools is likely to play an important role in the emerging assessment ecosystem, but, at present, the diverging views hinder coordination. This stifles more general progress toward the comprehensive approaches to assessing capabilities, risks, and potential harms necessary for creating effective standards and regulatory measures. Moreover, the divergence undermines practical research on methods relevant to effectively governing the rapidly emerging foundation model paradigm. This paradigm appears to still be in a pre-paradigmatic state (Kuhn 1962), where competing approaches are necessary. While caution may be prudent concerning foundation model regulation (Guha et al. 2023),

given the pace of progress, some steps will be necessary. Thus, coordination may soon be necessary to establish minimal yet robust assessment to support regulatory measures.

Second, the institutions, resources, and infrastructure needed to support federal agencies and regulators, academic researchers, external researchers, and independent third-party organizations, both for-profit and nonprofit, are critical to a healthy and thriving assessment ecosystem. The federal government has been working to establish institutions and resources, such as the NIST's U.S. AI Safety Institute (AISI) and the U.S. AISI Consortium, NIST's AI Risk Management Framework (NIST 2023), and the National AI Research Resource (NAIRR; (NAIRR 2023)). However, while these institutions and resources are steps in the right direction, tremendous work remains to ensure that they are integrated as effective elements of the AI assessment ecosystem—Executive Order 14110 (EO14110 2023) does not have the authority of legislation. Moreover, there needs to be sufficient funding allocated for these and additional efforts. Further, it is essential that potential regulatory efforts, new institutions, and new programs, including NIST's efforts, like the AISI and the AISI Consortium, or efforts like the NAIRR, support innovation and growth of the assessment ecosystem, as the current community working on this essential issue is unlikely to be able to keep up with growing demand.

Finally, as the foundation model paradigm is still pre-paradigmatic (Kuhn 1997), it is crucial to the health of the assessment ecosystem to identify salient questions for prioritizing research to support effective regulation without waiting until a new paradigm is established. Thus, in addition to research specific to the different schools described above, it is important to also prioritize more practical research questions that can help to guide regulators in assessment efforts to aid in mitigating societal-scale harms and risks (Chan et al. 2023; Weidinger et al. 2023).

Steps Toward a Healthy Ecosystem

We have identified three challenges to the flourishing of a healthy assessment ecosystem for foundation models: divergent perspectives among stakeholders, the need for strong institutions and ample resources to support growth, and the need to prioritize practical research questions to aid regulators and other stakeholders in better navigating the rapidly evolving assessment ecosystem. Below, we recommend steps for grappling with these issues and pushing the ecosystem in a direction conducive to flourishing.

It is critical to take steps to reconcile diverging perspectives, which, while not inherently bad, are inhibiting constructive dialogue on model assessment between some of the different schools (NIST 2024). We propose that NIST establish a new working group or task force to create a list of key definitions to facilitate more effective communication among stakeholders. We suggest that this working group be asked to consider the full space of future foundation model capabilities, including across all stages of a model's life cycle—i.e., pre-training, fine-tuning, and post-deployment enhancements. We also recommend that NIST require this task force/working group to elicit stakeholder feedback and create a vision and a roadmap for foundation model assessment.

To establish the infrastructure and resources for a thriving ecosystem, we look toward legislators. While EO 14110 appears to be a step in the right direction, it is important that legislators take prompt and informed yet cautious action. Moreover, lawmakers need to work with various stakeholders in the AI assessment community to ensure that NIST, the new AISI, the AISI Consortium, the NAIRR, and any other new institutions that are established as a result of legislation are properly prioritized in appropriations. We feel it would be especially valuable to establish a National Center of Excellence (CoE) for AI Safety and Assessment. We envision this CoE as housing an ultra-secure cluster as part of the NAIRR—this would go beyond the secure level proposed for the NAIRR, e.g., being air-gapped—in order to ensure AI developers that their state-of-the-art model weights could be stored for research purposes, safe from exfiltration. This would provide academic researchers with an unprecedented opportunity to conduct safety and assessment research requiring full model access (e.g., mechanistic interpretability research) on proprietary models. This could also help to foster growth of the ecosystem by providing grants to academics and for-profit firms to use the ultra-secure cluster, establishing a flagship research hub to further facilitate coordination and collaboration within the community. Additionally, the cluster could be dual-use, to support regulatory or other high-securiy government assessment needs. Lastly, we feel that it is essential that potential regulatory efforts and new institutions support innovation and growth of the assessment ecosystem; we feel it especially critical to foster investment and growth of for-profit assessment firms and to aggressively support academic assessment research. While our suggestions have been U.S.-centric, other stakeholders, such as the U.K. AI Safety Institute, may also play a constructive role in a healthy, thriving ecosystem.

Finally, we look to practical research. Here, we provide a few examples, but the list is non-exhaustive, and we recommend a research agenda on the topic be prioritized. One question is how to address challenges posed by fine-tuning away model safety and content restrictions, something particularly pertinent for ensuring a vibrant open source community. It is also especially relevant to determining the degree of regulatory requirements for fine-tuned models, and could provide forewarning to the assessment community of increased demand or the need for alternative approaches (e.g., Scale). Another underexplored area is continual post-deployment model assessment; Scale and others have proposed automating the process using other foundation models (ScaleAI 2023), but more research will be necessary. Yet another area of research is anticipatory assessment of models for potential risks or emergent capabilities that may arise from post-deployment enhancements.

In conclusion, while we strongly support these recommendations and feel they will greatly benefit the ecosystem, we encourage criticism and alternative ideas. Therefore, we hope that this extended abstract can serve as a catalyst for discussions about mid-to-long-term visions of what a healthy and thriving assessment ecosystem may look like.

Acknowledgements

This was drawn from a working manuscript to be submitted to NIST in response to a request for information on safe, secure, and trustworthy AI development.

References

Arcas, Y.; and Norvig. 2023. Artificial General Intelligence is already here.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krasheninnikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.

Dentella, V.; Murphy, E.; Marcus, G.; and Leivada, E. 2023. Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv preprint arXiv:2302.12313*.

EO14110. 2023. Executive Order 14110 (U.S.) 2023. Safe, Secure,224 and Trustworthy Development and Use of Artificial Intel-225 ligence.

Gruetzemacher, R.; and Whittlestone, J. 2022. The transformative potential of artificial intelligence. *Futures*, 135: 102884.

Guha, N.; Lawrence, C.; Gailmard, L. A.; Rodolfa, K.; Surani, F.; Bommasani, R.; Raji, I.; Cuéllar, M.-F.; Honigsberg, C.; Liang, P.; et al. 2023. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review, Forthcoming*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jain, S.; Kirk, R.; Lubana, E. S.; Dick, R. P.; Tanaka, H.; Grefenstette, E.; Rocktäschel, T.; and Krueger, D. S. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*.

Kinniment, M.; Sato, L. J. K.; Du, H.; Goodrich, B.; Hasin, M.; Chan, L.; Miles, L. H.; Lin, T. R.; Wijk, H.; Burget, J.; et al. 2023. Evaluating language-model agents on realistic autonomous tasks. *arXiv preprint arXiv:2312.11671*.

Kuhn, T. S. 1997. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

NAIRR. 2023. Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem An Implementation Plan for a National Artificial Intelligence Research Resource.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

NIST. 2024. NIST Secure Software Development Framework for Generative AI and for Dual Use Foundation Models Virtual Workshop.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

ScaleAI. 2023. Test and Evaluation Vision.

Sharkey, L.; Ghuidhir, C. N.; Braun, D.; Scheurer, J.; Balesni, M.; Bushnaq, L.; Stix, C.; and Hobbhahn, M. 2024. A Causal Framework for AI Regulation and Auditing.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.

Zhou, L.; Moreno-Casares, P. A.; Martínez-Plumed, F.; Burden, J.; Burnell, R.; Cheke, L.; Ferri, C.; Marcoci, A.; Mehrbakhsh, B.; Moros-Daval, Y.; et al. 2023. Predictable Artificial Intelligence. *arXiv preprint arXiv:2310.06167*.