Envisioning a Thriving Ecosystem for Testing & Evaluating Advanced AI

Ross Gruetzemacher[1,2,3], Kyle Kilian[2,4], Iyngkarran Kumar[2], David Manheim[5,6], Mark Bailey[7], Pierre Harter[1], José Hernández-Orallo[8,3,9]

1 Wichita State University, 2 Transformative Futures Institute, 3 Centre for the Study of Existential Risk, University of Cambridge, 4 Center for the Future Mind, Florida Atlantic University, 5 Technion, Israel Institute of Technology, 6 Association for Long Term Existence and Resilience, 7 National Intelligence University, 8 Universitat Politècnica de València, 9 Leverhulme Centre for the Future of Intelligence, University of Cambridge

Abstract

We are witnessing the dawn of an era of advanced AI systems as the space of foundation models' capabilities continues to grow with ever more computational resources and increasingly higher-quality data being allocated for training. These systems have great potential to increase economic productivity and lead to new scientific discoveries to the benefit of all. At the same time, these systems pose novel risks, and in constituting a new paradigm, they have sparked the formation of four unique schools of thought concerning testing and evaluation of AI systems' risks and capabilities. These four schools include: 1) the testing, evaluation, verification, and validation (TEVV) school, 2) the benchmark school, 3) the 'evals' school (short for evaluations), and 4) the cognitive school. Drawing from our initial assessment of the testing and evaluation ecosystem, we identify three central obstacles impeding progress toward a healthy and vibrant testing and evaluation ecosystem: fragmentation of the ecosystem amidst the distinctive schools of thought, insufficient infrastructure and resources, and a number of pressing, practical research questions that are currently overlooked. Overcoming these obstacles will require first unifying the testing and evaluation ecosystem by establishing a shared understanding and vision for the future, engineering an environment aimed at fostering growth, and identifying and resolving salient research questions relevant to guiding legislators and regulators in establishing and growing the ecosystem. We provide background for understanding the current status quo; we describe the different schools of thought; we describe the central challenges to the health of the ecosystem; we propose a vision for a healthy and thriving ecosystem as well as potential steps to help in navigating the challenges and obstacles that may stand in the way. In conclusion, we recommend a list of salient action items for NIST and legislators.

Challenges for the Current Testing & Evaluation Ecosystem

Foundation models are powering increasingly advanced AI systems which differ from more traditional, narrow AI systems in that they can be easily trained to perform a broad range of tasks with minimal domain-specific fine-tuning. This characteristic suggests many economically valuable applications and is expected to drive economic growth and advance scientific progress over the coming decades. However, this characteristic also creates challenges for mitigating the harmful consequences of this emerging technology.

To mitigate the societal harms from AI systems, testing and evaluation (T&E) must be employed to ensure that deployed systems are used in a safe and responsible manner. Foundation models are considered to be a new paradigm in AI research, and by virtue of the fact that they exhibit capabilities such as reasoning and planning that enable them to perform well across numerous consequential domains, it is evident that they require fundamentally different approaches to T&E. Due to such challenges, significant effort is being made to develop techniques for testing, evaluation, and red teaming of capabilities of state-of-the-art systems.

We see three critical challenges that must be navigated to ensure a healthy and thriving T&E ecosystem for AI. First, research on foundation models is in a preparadigmatic state (Kuhn 1962), and as a result the T&E ecosystem is and will continue to be fragmented into competing schools until these schools either merge or one school prevails; this process is natural and critical to scientific progress but is especially problematic for regulatory efforts. Second, there is the need for well-resourced institutions and infrastructure to support research and regulatory efforts. Third, there is a need to conduct practical research at the intersection of the different schools in order to help regulators and members of the T&E community successfully navigate these challenges.

Fragmentation of the testing and evaluation ecosystem: We identify four significant schools of thought in the AI T&E ecosystem: We refer to the first of these as the testing, evaluation, verification, and validation (TEVV) school, as this is a rigorous and robust process closely associated with NIST and other similar government institutions. TEVV extends beyond AI systems to other realms of engineering (e.g., software, robotics) where the primary objective is certifying systems utilizing metrology. However, new and different schools are gaining interest in the T&E ecosystem. For example, recent years have seen increasingly rapid progress on benchmarks, an ever-evolving subdomain of machine learning aimed at assessing progress; interest here has been especially pronounced for language models. Beyond these well-established research areas are emerging schools of thought that relate more specifically to the new paradigm of foundation models. The most dominant of these—the ‘evals’ school—derives from the cybersecurity community and involves red teaming and ‘evals’, an umbrella term short for evaluations that is often used

to imply a specialized class of testing intended to stress test foundation models by identifying risks from unanticipated but potentially dangerous emergent capabilities of systems at the frontier of AI research (i.e., frontier AI systems). The final school of thought derives from cognitive psychology and focuses on separating capabilities from risks and identifying them in a more fundamental way (e.g., psychometrics); however, these approaches can sometimes be thought of as falling under the 'evals' umbrella, as well. Table I further described the four central schools comprising the AI T&E ecosystem.

Table I

School	Description
TEVV	This is the established school of thought on testing, evaluation, verification, and validation; it is exemplified by governmental agencies and laboratories working on metrology, such as NIST (e.g., the AI RMF 1.0), the European Commission's initiatives on AI Testing and Experimentation Facilities (TEFs) and associated Testing Sandboxes, and the Laboratoire National de Métrologie et d'Essais (LNE) in France.
Benchmarks	Exemplified by the creation or compilation of datasets that can be used to evaluate model performance, with the goal of identifying the best models, either in a specific domain or more generally; it is common in machine learning competitions and involves aggregated metrics of performance on the task(s). This approach is epitomized by leaderboards in competitions like those collected in paperswithcode.com.
Evals	This school is the newest perspective and the one most closely aligned to AI developers but also the one furthest from traditional TEVV practices; it is typically associated with extreme or harmful capabilities "evaluations" (Shevlane et al. 2023) and organizations focusing on these risks such as leading AI developers (e.g., Meta, Google DeepMind, OpenAI, Anthropic) and third-party auditors (e.g., Apollo Research, METR).
Cognitive	Exemplified by perspectives that come from the behavioral sciences. Here, capabilities are separated as latent traits that explain and predict performance. With the advent of foundation models and increasingly capable AI systems, the need for capability-oriented evaluation rather than task-oriented evaluation (Hernandez-Orallo 2017a; 2017b) has become a priority, because we cannot anticipate the full range of tasks users are going to ask of these systems. Also there is increasing interest from psychology to use or adapt traditional methodologies and techniques, such as psychometrics, to evaluate these systems (Wang et al. 2023, Ivanova 2023).

Some groups are working to mix methods from different schools, but there are many challenges. Work mixing and collaborating in areas of overlap between the schools would be

beneficial to the broader T&E ecosystem, but even communication is difficult concerning the nature and problems associated with the new foundation model paradigm.

Insufficient infrastructure and resources: Also critical to the health and vitality of the T&E ecosystem are the institutions, resources, and infrastructure available to support regulatory efforts and external research, as well as a suitable political and regulatory climate to support growth and investment in the community. Without a conducive environment for growth and flourishing, the T&E ecosystem is unlikely to realize its potential. Thus, it will be necessary to ensure that legislation and appropriations are very well thought through to create favorable conditions for the ecosystem's growth for all types of stakeholders—in academia, government, and in the private sector.

While the U.S. federal government has been working to establish institutions and resources, such as the NIST's USAISI and the USAISI Consortium, NIST's AI RMF (NIST 2023), and the National AI Research Resource (NAIRR; Prashar et al. 2023), there is likely a significant role for the private sector to play (Scale 2023). AI developers as well as other well-capitalized firms have significant soft power that could be used to establish industry bodies that could play a role in the T&E ecosystem. Additionally, private equity financiers could be able to play a constructive role if legislation provides favorable conditions for establishing new and lucrative markets for AI auditing—and AI safety—services.

Underexplored practical research questions: It is critical that practical research questions relevant to regulatory efforts and to practically managing T&E efforts are prioritized. For example, due to the ability of foundation models to be continually trained via fine-tuning, deployed AI systems or models being fine-tuned for domain-specific applications could unlearn safety or content restrictions put in place during the initial T&E stage. Without well-regulated and continual monitoring, this could occur without the knowledge of AI developers as they update their deployed systems. However, much is unclear about how to manage an environment where most systems are able to unlearn safety or content restrictions, and it is unclear how best to manage a continual T&E process. Thus, while it is necessary for research in each of the schools to proceed vigorously, it is also essential to prioritize this research of practical concern to regulators and the ecosystem itself.

Other risks: There are a variety of other risks we mention briefly. Politicization of AI risks—such as a politicization of near- versus long-term risks (Sætra and Danaher 2023)—could prove detrimental to efforts to foster growth and flourishing of the T&E community. This problem could result from actions of Congress or through politically motivated Executive management of federal agencies charged with critical T&E tasks such as the USAISI. Another serious concern is the role that international relations could play involving future advanced AI systems; it is imperative to coordinate internationally and to think of the T&E ecosystem as an inclusive global community as accident and structural risks are shared collectively by the entire international community (Gruetzemacher et al.

2023). Other risks involve the potential for T&E exemptions on the basis of national security. It will be imperative to take steps to ensure that military and intelligence agencies are actively engaged in the community.

Envisioning a Thriving Ecosystem

We envision a thriving T&E ecosystem for advanced AI as one that is able to accomplish several core functions to address the three core challenges described. Doing this will require a robust foundation of agencies and institutions to coordinate efforts among stakeholders and to take on leadership roles; these include NIST, the NSF, the NAIRR, the U.S. AI Safety Institute (USAISI), the U.S. AI Safety Institute Consortium, and the establishment of a National Center(s) of Excellence in AI Safety, Testing, and Evaluation. NIST, in its role leading the USAISI and USAISI Consortium, should take steps to encourage more constructive dialogue among the different schools, and should encourage establishing a practical research agenda to help regulators and the T&E community better understand and navigate the unique practical challenges for T&E of foundation models. Moreover, a healthy and thriving T&E ecosystem requires quick action to grow the number of evaluators and other stakeholders (e.g., staff in government labs, academics, for-profit and nonprofit evaluators and safety research firms) as well as extensive collaboration and discussion among the various stakeholders. Therefore, the vision presented here is not intended to be a solution, but rather a launching point for inclusive discussions. Below we suggest goals and concrete action items that we feel to be objectively constructive and urgent.

Action Items:

We list the goals and proposed action items below. Numbers 1.A, 1.B, 2.C, 3.A, 3.B, and 3.C are specific or relevant to NIST. Numbers 2.A and 2.B provide recommendations for lawmakers to support NIST and the T&E ecosystem.

1. **Goals: support continued work across all schools, encourage coordination between schools for research of practical concerns to support NIST's USAISI**
    A. **Action item:** NIST establish a task force to create list of definitions of key terms in order to facilitate more constructive discussions and to reduce dissonance between the different schools of thought
    B. **Action item:** NIST establish a task force to elicit and aggregate stakeholders' opinions to create a vision and corresponding roadmap for the USAISI

2. **Goals: establish and strengthen the necessary institutions and infrastructure to support a healthy and thriving T&E ecosystem**

A. **Action item:** legislators provide sufficient funding for infrastructure, resources, and both new and existing institutions supporting the AI T&E ecosystem, especially NIST, the USAISI, and the USAISI Consortium
B. **Action item:** legislators utilize available levers at their disposal to encourage growth and innovation in the T&E ecosystem through a combination of public/private investment
C. **Action item:** legislators, NIST, and the NSF coordinate to establish a National Center of Excellence (CoE) in AI Safety, Testing, and Evaluation to house a dual-use HPC resource to support the NAIRR and NIST; form a research hub at the facility to coordinate, grow, and energize the T&E ecosystem

3. **Goals: identify potential practical challenges for the T&E ecosystem, and prioritize research to help navigating them:**
   A. **Action item:** NIST's USAISI strongly encourage research on T&E requirements over the entire foundation model life cycle, e.g., pretraining, fine-tuning, post-deployment enhancement, as well as other research of urgent practical concern, e.g.:
      a. Research related to the the frailty of fine-tuning (Jain et al. 2024) that has significant and direct implications for T&E standards
      b. Research related to continual monitoring, testing, and evaluation of models post-deployment as this has significant and direct impcliations for T&E standards
   B. **Action item:** NIST's USAISI establish a task force or working group to create a research agenda, prioritizing open practical questions tied to standards and regulation, with a particular focus on practical regulaory and governance issues related to frontier AI systems

Action Items Discussed

Unification of the Ecosystem

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. In the NIST SSDF for Generative AI[1] different invited guests used different terms, with some using more specific terms to precisely articulate concerns related to foundation models. Moreover, the 'evals' school uses the term evals in lieu of evaluations, but the semantics ascribed to it are not consistent with definitions of evaluations used in other machine learning contexts (Raschka 2018; Japkowicz 2006). Evals more practically refer to adversarial testing and red teaming approaches, but this is likely not obvious to

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

some from other schools involved in NIST's USAISI Consortium. Such a lack of a shared understanding can make effective and constructive discussions very difficult concerning wicked problems[2] (Conklin 2007), and is especially important in software engineering contexts when preventing a negative outcome is essential (Glinz and Fricker 2015).

Action Item 1.A: NIST establishes a 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. Considering the need for the T&R ecosystem to grow and expand, terms with well-established existing definitions should continue to be used in a way that is not confusing to domain experts that are interested in contributing. NIST could set guidelines for the task force to follow, and while participation in the task force should be limited to members of the USAISI Consortium who volunteer to partake, NIST should strive to ensure that each of the schools are well-represented.

This document was created with the intention of trying to spark discussions about longer-term visions for the T&E ecosystem. NIST is correctly focused on the heavy load that they were tasked with by EO 14110, and those working directly on evaluations are focusing entirely on understanding and innovating adversarial testing and red teaming techniques for stress testing the next generation of highly capable foundation models. Therefore, there appears to be a gap in the current work with respect to planning for the future.

Action Item 1.B: NIST establishes a task force or working group to elicit feedback from all key stakeholders in order to create a vision of and a roadmap toward a healthy and thriving T&E ecosystem. The discussion of the ecosystem, and the creation of such a document, could go a long way in providing the sort of guidance that the ecosystem currently lacks, but which could be very effective at precipitating growth and coherence of the divergent schools. A single unifying document, including definitions created from the above proposal, as well as a vision and road map for the T&E ecosystem, may be most useful. However, a document specifying relevant definitions alone, or to supplement another document would be of significant value in its own right.

Strengthening Institutions, Infrastructure, and Resources

To establish the infrastructure and resources for a thriving ecosystem, we look toward legislators, who are in an ideal position to influence the institutions, infrastructure, and resources needed to support a thriving T&E ecosystem. While EO 14110 appears to be a step in the right direction, it is important that legislators take prompt and informed yet

[2] AI governance and AI alignment are thought to be super wicked problems (Gruetzemacher 2018), a class of coordination problems more challenging than wicked problems. Other examples would be Covid19 or climate change (Auld et al. 2021).

cautious action to build on the initiatives and tasks that the EO assigned to NIST. Moreover, lawmakers need to work with various stakeholders in the AI assessment community to ensure that NIST, the new USAISI, the USAISI Consortium, the NAIRR, and any other new institutions that are established as a result of legislation are properly prioritized in appropriations.

Action Item 2.A: Legislators prioritize appropriations to NIST and the USAISI, as well as efforts like the NAIRR, or any potentially new institutions, resources, or infrastructure established by legislation for the purpose of supporting T&E and mitigating risks from advanced AI, in the manner that other concerns of national security are prioritized in federal appropriations. It is critical that appropriations for NIST be sufficient to launch and recruit talent to the USAISI, and given recent funding shortfalls at NIST, sufficient funding for other operations at NIST would be prudent. Anthropic has recommended $15M for NIST, and an additional $25M toward addressing issues with crumbling infrastructure, such as for maintaining laboratory facilities, would be a good start.

Action Item 2.B: Legislators should use all levers at their disposal, possibly including regulatory requirements, legal responsibilities, required reporting, tax incentives, and/or subsidies, to encourage private investment to establish and grow this new sector, as strong private sector work in this area will be positively correlated with a healthy and thriving T&E ecosystem. There is a large variety of actions that legislators could take to foster growth in this area, such as through allowing a limited number of low-interest or no-interest startup loans for qualified applicants interested in entering the sector. They can also provide grants to support R&D activities of existing private auditors or evaluators; monetary grants or computational allocations of NAIRR resources are both potentially appealing incentives. Establishing tiers of evaluations based on risk levels and system size (Gruetzemacher et al. 2023) could help reduce demand for the most labor intensive testing low, while also enabling less intensive testing to be required of a larger number of models, providing more opportunities for new entrants to the market. Tiers would be dynamic, tracking things like AI safety levels. In addition, exemptions to the H-1B visa cap could be made for employees at designated T&E firms; even per-country cap exemptions could be helpful.

Action Item 2.C: The ecosystem needs a pillar or pillars to serve as a foundation; to these ends we propose NIST, the NSF, and legislators work together toward establishing a National Center(s) of Excellence (CoE) in AI Safety, Testing, and Evaluation. This would not be the monolithic center for safety, testing, and evaluation, but would instead be a leader and pioneer in a broader T&E ecosystem; it would play a significant role in testing, evaluations, and alignment research; and it would play a critical role in ensuring the timely development of standards. To enable this, the hub would need both technical expertise and computing resources.

In the CoE's core facility we envision a dual-purpose HPC cluster designed in a cost-efficient manner, initially to prioritize inference on foundation models, especially those pushing the frontier of state-of-the-art AI capabilities. Foremost, this AI cluster would serve as part of the NAIRR dedicated specifically to research on AI safety, testing, and evaluation. Specifically, this research would focus on approaches requiring full model access, like mechanistic interpretability, using the AI developers' state-of-the-art systems (the cluster would be ultra-secure—described below—to ensure the safety of AI developers' models for research). The CoE cluster would also serve as a facility with the ability to be used for any needs that may arise for NIST and the USAISI, such as pre-deployment testing or verification of systems test results, or for other federal agencies T&E needs that would necessitate such an extreme degree of security, such as might be the case for advanced national security related AI systems.[3]

It will be essential that the AI cluster of the CoE be *ultra-secure*; this would require the cluster to be air gapped and to generally adhere to protocols associated with the sensitive compartmented information facilities (SCIFs) used by intelligence agencies and government officials when handling highly classified information. We use the term ultra-secure to refer to this level of security in order to distinguish it from the NAIRR secure notion (NAIRR Task Force 2023).

There are numerous reasons for security of such an extreme degree; here we give three examples. Most importantly, in order for labs to trust the use of the facility for storing their model weights, either for pre-deployment testing or verification purposes, or for research purposes, SCIF-like information security will be necessary to ensure that this valuable and strategic proprietary information is not susceptible to exfiltration; by default, no digital information would be directly transferred from the ultra-secure facility, and any reports and results from research would be required to be printed on paper. The facility could explore options for relaxing the extremely strict security requirements; for example, a model that allowed tests or evaluation scripts to be imported to the cluster, allowing outside researchers to run the equivalent of read-only queries, could be explored, but use of any such models would need approval from all AI developers sharing model weights and data. The facility could be used for research on what information security measures are necessary for different AI safety levels, and could act as a model for the robust security measures necessary for training the most capable frontier AI systems.

Additionally, rigid information security protocols would enable a CoE's AI cluster to be used for highly sensitive evaluations or pre-deployment testing. This could include use in the testing of intelligence or defense systems, or, for testing and evaluation of dual-use AI technologies (e.g., Mouton et al. 2023). Further, the cluster would be used for other sensitive

[3] To use a NAIRR resource for government purposes might require a new tier of use be added to the NAIRR Bill; a new tier for government use would enable the cluster to be designated in the dual-purpose fashion proposed.

evaluations, such as for autonomous replication and adaptation (ARA; Kinnement et al. 2023) where it would be important to keep models exhibiting dangerous emergent capabilities from proliferating. Because the secure cluster would not be afforded the ability to take real-world actions, there will be a need for high-fidelity simulated environments for pre-deployment testing,[4] which will need to become a research priority of the ecosystem.

Institutions housing a CoE could charge AI developers fees to conduct any pre-deployment tests or verification tasks required for meeting standards set by NIST or NIST's AI Safety Institute. Other federal agencies could also pay fees for their use of the CoE's AI cluster. These fees could go to cover personnel costs involved in creating the pre-deployment tests, as well as to support the HPC resources utilized during the testing. We note that the AI CREATE Act, or other relevant legislation, would need to establish a use tier enabling federal agencies to take advantage of such a fee structure.

If labs were comfortable with the information security levels guaranteed by the CoE, they would be comfortable sharing their deployed state-of-the-art model weights or training data[5] at the facility. Given such access, this facility would be incredibly beneficial to the T&E ecosystem for fostering research because there would be no other place where research on state-of-the-art models involving full model access or other sensitive information like training data could be conducted for those outside of the AI developers labs. Such research would benefit not only the academic community, but also the AI developers, providing them with additional talent and ideas to advance AI safety, testing, and evaluations research. The mutually beneficial nature of this arrangement would be very hospitable to the growth of the T&E ecosystem. Confidentiality agreements would be required, and strict rules would be necessary for publication, if deemed permissible by the CoE organizers and the AI developers; however.

The need for the CoE HPC resource to be air gapped would require NAIRR users of the cluster to be located in a central location physically close to it so that they could enter the facility to conduct experiments. Thus, the CoE would become a geographical hub for AI safety, testing, and evaluations research. Creating physical hubs of talent in this niche domain could significantly help to foster robust and rapid growth of the research community working on problems related to AI safety and T&E for advanced AI. Organizations would not necessarily need to relocate a large number of employees to the hub, but it would be necessary to establish small remote offices of scientists that would be tasked with conducting experiments on the secure with a hub of remote offices for companies and other organizations seeking to use this component of the NAIRR. Thus, the CoE may need to be in a research park or area with robust physical security but also with office space for hosting firms and researchers intent on utilizing the CoE's resources. In

[4] Sharkey et al. (2024) suggest that auditing be necessary for a staged deployment process, providing increasing affordances to the system as a result of successful audits. Pre-deployment testing could be required prior to frontier systems being afforded critical abilities to take actions in the real world (e.g., Internet access) as opposed to a simulated environment.
[5] Another valuable and unique role the facility could play would be to provide a safe environment for auditing training data.

planning for the CoE, leaders could look to other models for inspiration, such as the European Commission's European Digital Innovation Hubs.[6]

Initially, the CoE could pilot a smaller cluster of between 256 and 512 memory-efficient GPUs.[7] If utilization of the secure cluster was able to grow to acceptable rates for research, then future scaling to a larger cluster could be justified. A pilot period would also provide NIST and other federal agencies with time to better understand their needs for such a resource.

Salient Practical Research Directions

At present, the most capable foundation models undergo various forms of fine-tuning prior to deployment in order to curb undesirable behaviors exhibited after instruction fine-tuning of the original pretrained model. However, numerous recent studies have highlighted the frailty of current safety fine-tuning methods (Jain et al. 2024; Qi et al. 2023). For example Jain et al. show that fine-tuning does not fundamentally alter a model's behaviour but instead adds a superficial 'wrapper' on top of the existing capabilities, leading to an illusion of modified capabilities. As a result, a small amount of further fine-tuning can recover the undesirable behaviour that safety fine-tuning was supposed to have removed. Lermen et al. (2023) corroborates this, removing safeguards from Llama (Touvron et al. 2023) with minimal compute.

If the robustness of safety fine-tuning methods does not improve, strategies like the full release of model weights—as is the case for open source models—appear prone to misuse, and even fine-tuning APIs such as that provided by OpenAI[8] should be carefully scrutinized to ensure that safety guardrails are not removed.

Action Item 3.A.a: NIST forms a task force or working group to identify and prioritize research questions that shed light on the limitations of overreliance on fine-tuning as a means of ensuring AI systems' safety. Scoping this problem and identifying the critical questions necessary for determining regulatory action is critical and urgent. Such work could shed light on whether developing safety fine-tuning methods that alter a model's capabilities at a far deeper level than current methods is a feasible near-term solution, or whether the T&E ecosystem needs to develop standards and protocols to manage the issue. Scale (2023) envisions this necessitating so much T&E of fine-tuned models as to require automation of T&E via foundation models; if this is in fact required, it is imperative to inform legislators and regulators so as to take the necessary actions.

[6] See https://digital-strategy.ec.europa.eu/en/activities/edihs.
[7] We anticipate a CoE cluster to be designed to utilize GPUs that provide the most co-processor memory per dollar when considering all other factors. We anticipate somewhere between 36TB and 48TB being ideal.
[8] See https://platform.openai.com/docs/guides/fine-tuning.

Continual monitoring of a model post-deployment will also be critical to any successful T&E regime. Once deployed models will encounter and be required to operate in environments that were not encountered during training and T&E. Inevitably, models will encounter out-of-distribution scenarios precipitating unanticipated novel failure modes. Additionally, training of systems at the frontier of AI research will continue to require ever more computational resources. Thus, AI developers, once having invested millions or billions of dollars into training a system, will seek to utilize fine-tuning or RLHF to continue to enhance performance as much as possible, which becomes problematic for reasons discussed above. Alternatively, it is possible that systems are subjected to post-deployment enhancement, such as through the use of agentic wrappers or scaffolding, either by AI developers or independent actors. It could be particularly challenging if such enhancement were to be implemented in ways that had not been anticipated during pre-deployment T&E. Taken together, these examples underscore the necessity of a variety of fundamentally different approaches to continual post-deployment monitoring.

Action Item 3.A.b: NIST forms a task force or working group to map the space of risks over AI systems entire life-cycles, and identify the salient and open research questions concerning such risks most relevant to guiding legislators and regulators. The findings could be disseminated, and prioritized by the USAISI Consortium members or a National CoE for AI Safety, Testing, and Evaluations. Post-deployment enhancements are broad in scope, but the entire space of risks over AI systems' life-cycles would need to also include sociotechnical AI safety concerns (Lazar and Nelson 2023; Weidinger et al. 2023).

The two proposals we have made in this section relate to research streams with obvious implications on policy decisions that are currently understudied and poorly understood. There are likely numerous more similarly essential research directions worth exploring, but the T&E ecosystem is still nascent and fragmented, and there are limited or no resources available able to provide a succinct overview of the ecosystem, or a comprehensive review of the various active research directions or projects for even one of the four schools. We have proposed that NIST take action to create a vision and a roadmap for the T&E ecosystem, which would be valuable to existing members of the community as well as new parties interested in contributing. However, a document clearly identifying the many very important research directions of direct and practical relevance to policymakers could be very beneficial for recruiting interested members of the community and for ensuring that new members of the community could quickly join and make substantive contributions.

Action Item 3.B: NIST creates a task force or working group to create a broad and comprehensive research agenda for T&E research of a practical nature. The two previous proposals in this section seem particularly urgent, but could be included here if a single mandate is more tractable given the resources and bandwidth of the USAISI Consortium. Aside from the previous proposals, there are many directions that a research

agenda for the T&E ecosystem would need to explore. Of immediate practical relevance might be trying to identify how best existing work from the different schools might be integrated. Regardless, it seems self-evident that a peer-reviewed and collaborative research agenda, even if only at a high-level, and lacking in granularity, would immediately be an asset to the T&E ecosystem and move the needle in the right direction.

Key Takeaways

We have presented several concrete suggestions. We highlight key takeaways:

- The most unique action item suggested is 2.C, and we urge the NIST USAISI to explore how this could be beneficial both to the community and to tasks NIST has been charged with in EO 14110.
- Action Items 1.A and 1.B are necessary to unify the T&E ecosystem, which is critical to NIST receiving more effective support from the USAISI Consortium.
- Action Items 3.A and 3.B are also critical, and require NIST's leadership to unite the T&E ecosystem around a common research agenda focused on answering the practical questions necessary to support NIST and legislators in creating legislation and standards necessary to navigate the emergences of this extremely powerful new technology.

Acknowledgements

We thank Alan Chan, Stephen Casper, and Herbie Bradley for their comments at different stages of this, on this document and the full manuscript.

References

Auld, G., et al. 2021. Managing pandemics as super wicked problems: lessons from COVID-19 and climate. Policy sciences, 54.

Conklin, J., 2005. Dialogue mapping: Building shared understanding of wicked problems. John Wiley & Sons, Inc..

Glinz & Fricker. 2015. On shared understanding in software engineering. CS-Research & Development, 30, pp.363-376.

Gruetzemacher, R., 2018, December. Rethinking AI strategy & policy as entangled super wicked problems. AIES (pp. 122-122).

Gruetzemacher, R., Chan, A., Frazier, K., Manning, C., Los, Š., Fox, J., Hernández-Orallo, J., Burden, J., Franklin, M., Ghuidhir, C.N. and Bailey, M., 2023. An International Consortium for Evaluations of Societal-Scale Risks from Advanced AI. arXiv preprint arXiv:2310.14455.

Hernández-Orallo, J., 2017a. The measure of all minds: evaluating natural and artificial intelligence. Cam Uni Press.

Hernández-Orallo, J., 2017b. Evaluation in AI: from task-oriented to ability-oriented measurement. AI Review, 48.

Ivanova, A.A., 2023. Running cognitive evaluations on large language models: The do's and the don'ts. arXiv preprint arXiv:2312.01276.

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.

Japkowicz, N., 2006, Why question machine learning evaluation methods. AAAI workshop on evaluation methods for ML.

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

Kuhn, T.S., 2012. The structure of scientific revolutions. University of Chicago press.

Lermen, S., Rogers-Smith, C. and Ladish, J., 2023. LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. arXiv preprint arXiv:2310.20624.

MOUTON, C.A., LUCAS, C. and GUEST, E., The Operational Risks of AI in Large-Scale Biological Attacks.

NAIRR Task Force. 2023. Strengthening & Democratizing The U.S. AI Innovation Ecosystem. Congressional Report.

Parashar, M., DeBlanc-Knowles, T., Gianchandani, E. and Parker, L.E., 2023. Strengthening and democratizing artificial intelligence research and development. Computer, 56(11), pp.85-90.

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.

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808.

Sætra, H.S. and Danaher, J., 2022. To each technology its own ethics: The problem of ethical proliferation. Philosophy & Technology, 35(4), p.93.

Scale. 2023. Testing and Evaluation Vision. Scale, blog. https://scale.com/guides/test-and-evaluation-vision#vision-for-the-t&e-ecosystem

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. and Ho, L., 2023. Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Wang, X., Jiang, L., Hernandez-Orallo, J., Sun, L., Stillwell, D., Luo, F. and Xie, X., 2023. Evaluating General-Purpose AI with Psychometrics. arXiv preprint arXiv:2310.16379.