Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects

Anonymous Author(s)
Affiliation
Address
email

Abstract

Conventional AI evaluation approaches concentrated within the AI stack exhibit systemic limitations for exploring, navigating and resolving the human and societal factors that play out in real world deployment such as in education, finance, healthcare, and employment sectors. AI capability evaluations can capture detail about first-order effects, such as whether immediate system outputs are accurate, or contain toxic, biased or stereotypical content, but AI's second-order effects, i.e. any long-term outcomes and consequences that may result from AI use in the real world, have become a significant area of interest as the technology becomes embedded in our daily lives. These secondary effects can include shifts in user behavior, societal, cultural and economic ramifications, workforce transformations, and long-term downstream impacts that may result from a broad and growing set of risks. **This position paper argues that measuring the indirect and secondary effects of AI will require expansion beyond static, single-turn approaches conducted in silico to include testing paradigms that can capture what actually materializes when people use AI technology in context.** Specifically, we describe the need for data and methods that can facilitate contextual awareness and enable downstream interpretation and decision making about AI's secondary effects, and recommend requirements for a new ecosystem.

1 Introduction

As AI technologies have become mainstream, the number of tools for evaluating them have exploded within a highly active and competitive area of development and research. Measurement provides AI practitioners with the opportunity to test and learn whether and how the technology they build works once deployed [1] Evaluation enables interpretation of measurement results to place them into context. Metrology, the science of measurement, provides the methods and definitions of measurement that enable the evaluation of all measurement results, including for AI systems. Metrology provides the foundations for estimating measurement uncertainty that can incorporate multiple sources of random and systematic error.

AI testing and evaluation is currently conducted within a computational and machine learning (ML) frame, with few systematic methods to account for the complex human, organizational and societal factors that inter-relate with the design, development, deployment and use of these technologies. This socio-technical [2] framing of AI technology is currently difficult for ML practitioners to operationalize,

[1] Measurement: (1) Quantitative measurement is the act or process of assigning a number or category to an entity to describe an attribute of that entity. ISO/IEC 24765:2017 (2) Qualitative measurement is based on descriptive data such as through observations, interviews, focus groups, or open-ended text fields in surveys.

[2] The term "socio-technical systems" was coined in 1951 by Eric Trist and Ken Bamforth[101]to describe the dynamic ways workers interact with technological systems in industrial settings.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Table 1: Mapping evaluation approaches to effects measured and typical questions they answer.

Evaluation Approach	Type of Effects (order)	What it measures	Answers questions like
Benchmarking	1st	Performance of the model/system *in silico*.	1. How often can the AI system produce the most accurate or relevant answer? 2. What is the inference runtime? 3. Did the model produce human-aligned responses?
Testing & Evaluation	1st, 2nd	Performance of the model/system *in silico* *in vitro* and *in situ*.	1. Does text summarization provide value for users? 2. Given current performance and user needs, should we expect productivity gains if we deploy this technology? If so, where?
Verification & Validation	1st, 2nd	Performance of the model/system *in silico*, *in vitro* and *in situ*..	1. Does the AI system consistently generate video content per user specifications? 2. Does the AI system classify output according to vendor claims?
Program Evaluation	2nd, 3rd	Real-world efficacy and relevance *in vitro* and *in situ*.	1. Do AI assistants improve the quality of work? 2. How will AI-driven productivity gains transform different employment categories over time?

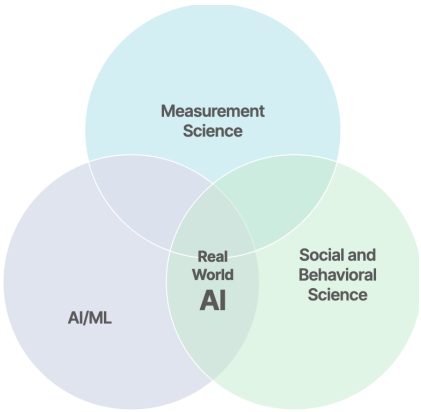
or to know where, when and how to include which types of contextual information across the technology lifecycle. **This paper argues that a new AI evaluation ecosystem is necessary to address current methodological gaps which impede the translation and contextualization of evaluation data and outcomes in the real world [107, 104, 16, 27, 29, 33, 85, 83, 35, 96, 37, 81, 76].** A real world AI evaluation ecosystem can enhance understanding of AI's second-order effects, drive the collection of datasets that are fit-for-purpose, foster innovation, and improve AI functionality.

1.1 The Measurement Challenge

The speed at which AI technology is advancing and being deployed and used across the globe [102] is not being met with equivalent evaluation paradigms for understanding its role and effects in societies. As a central topic of public policy efforts around the globe, questions about AI's secondary effects abound. Private industry, civil society, the public, and governments around the world are increasingly interested in how AI technologies will transform our culture, economy, workforce and the broader society[75, 74, 72, 73, 40, 43]. The current ecosystem to investigate these topics is fragmented, with no single evaluation toolbox or measurement infrastructure to account for AI's second order effects and place them into the broader context.

The predominant evaluation toolbox used by the ML research community, AI benchmarking, is designed to answer *first order* questions– about *what* AI systems can do based on direct measures of immediate system output. Another broad set of domains study AI's human, organizational and societal factors, which tend to focus more on second order questions such as the effects associated with how people leverage AI technology, and how and why those effects reverberate across society. Other fields can place these findings into context to forecast future technological and societal trends. Some approaches, like user simulations, can simultaneously model AI user behavior and evaluate system performance. The AI metrology community is also deeply engaged in the development of tools to assess systems in more realistic settings with the broader goal of ensuring AI system trustworthiness. This work includes development of definitions [10], and methods for calibration [106], and uncertainty quantification [39, 103] and propagation [100]. Yet, more effort is required, including efficient scalability and interpretation of measurement values.

Table 1 lists differences between the kinds of questions that can be answered by benchmarking, testing and evaluation, verification and validation, and program evaluation respectively. As methods



,

Figure 1: Disciplines at the intersection of Real World AI

move from first to second order and contextual detail increases, broader claims about AI technology become possible [1].

All evaluation methods have limitations which, when combined with the massive heterogeneity of how humans interact with AI in the real world, can present almost infinite complexities [71, 33] for building comprehensive paradigms. In addition to the technical and human factors of what and how to evaluate, there are numerous disciplinary and practical challenges to contend with. Reproducibility is an AI evaluation challenge of particular interest in computational domains and the social and behavioral sciences. In machine learning , reproducibility challenges include data leakage issues [51], non-systematic methods for curating training data, non-disclosure of training data information, unstable model versioning processes [50], and insufficient detail about experimental design, metadata, and related analytic processes[30, 80]. Advances in the culture of research practice have emerged to address the replicability crisis in the social and behavioral sciences [17]. Experiment pre-registration, open science standards, multiverse and sensitivity analyses, meta-analyses, and adversarial collaborations have led to varying levels of improvement [54, 6].

While a new ecosystem does not eliminate the above-listed challenges, a purpose-built community can concentrate on improving methods for evaluating second order effects. Figure 1 illustrates a real world AI community at the intersection of AI and ML, measurement science, and the social and behavioral sciences which can adapt and re-purpose methods, tools, metrics and practices to fuel deeper understanding of AI's complex societal challenges. This interdisciplinary community can collaboratively establish relevant measurement criteria, collect suitable datasets, formalize methods and practices and use resulting insights to produce better models for automation and real world forecasting and decision making.

2 AI Benchmarking

Model benchmarking is the de facto AI evaluation method. Benchmarking uses static datasets to assess performance of AI model capabilities on specific tasks at scale, often in comparison to humans. Evaluators use benchmark results to compare different models or systems on the same tasks. Benchmark suites are used to aggregate results and comprehensively assess capabilities, risks, and compliance. For example, systems may be tested for truthfulness [109], toxicity [42], and jailbreak vulnerability [24]. Benchmarking outcomes underpin AI system design, procurement, and oversight activities.

[1] In the context of AI evaluation, 1st-order effects are immediate system outputs, 2nd-order effects are longer-term impacts that may follow from system deployment, 3rd-order effects refer to broader changes that may result from AI's role in society. *In silico* refers to testing conducted via computational methods. *In situ* refers to observing a phenomenon in its natural location or context. *In vitro* refers to traditional laboratory experiments

2.1 The Practice of Benchmarking

The current benchmark landscape spans a wide range of tasks (text generation, question answering, summarization), modalities (text, code, audio, images, video), and evaluation dimensions such as factuality, fairness, safety, and alignment with human preferences[93, 14, 105, 109, 87, 63, 23, 53]. Recent benchmarks are built on various risk taxonomies to support an increased focus on AI risks and safety. For example, safety benchmarks can explicitly target risks posed by prompt injection[24] and data leakage [32], or be used to assess subtle failure modes that require multiple tests to capture nuanced system characteristics[82].

Benchmarks serve multiple purposes in the current evaluation landscape. They inform the setting of minimum performance thresholds-often through system requirements, regulatory norms, or technical standards but rarely define what constitutes “adequate” performance for a given policy context. With benchmarking defining the very notion of "success", these tests can shape perceptions of AI progress, influence research and development priorities, and inform investment cycles[35]. For this to be meaningful, benchmarking must be consistent and conducted before, during and after the development and deployment of AI tools and systems.

Benchmark design is complex but typically starts with identifying or acquiring curated datasets and specifying controlled tasks. Most generative AI benchmarks are conducted at the ‘single-turn’ level and assess independent query/response pairs[61]. Multi-turn benchmarks can be used to simulate more realistic dialogue and provide richer insights. Benchmarks rely on highly structured outputs such as multiple-choice, or short paragraph responses). Since there are technical limitations to which measures and characteristics can be analyzed at scale, evaluators often use LLMs to judge LLM benchmark output[110].

Benchmark test results are typically displayed via leaderboards, which provide a structured way to rank model performance and ease comparison [105, 89]. Undesirably, the evaluation community’s overreliance on leaderboards can lead to overfitting (models are optimized for test performance at the expense of real-world robustness) or to benchmark saturation, where further improvements on the test no longer translate to meaningful advances[99].

2.2 Selected Benchmarking Limitations for Real World Evaluation

Benchmarking’s prominence in AI evaluation has propelled the community to groundbreaking improvements and fostered global innovation but the outcomes can be limited. Benchmarking requires significant contextualization to serve the decision making needs of the many audiences interested in AI’s second order effects. AI benchmarks have been criticized for lacking internal and external validity [77, 64], encouraging leaderboard overfitting[78], and focusing narrowly on English-language tasks [70]. More broadly, they suffer from static design, lack of systematization, limited stakeholder involvement, and a failure to reflect cultural nuance. The static nature of benchmarking may obscure emergent behaviors, security vulnerabilities, or context-specific failures that only surface in deployment or over longer time periods. It is difficult to construct benchmarking tasks that naturally elicit generative AI risks – such as harmful bias, hallucinations, or user over-reliance – [83, 1, 86, 13, 49, 31, 25] and their associated impacts[77, 7], or the broad range of user responses and behavior that may arise from LLM-based personalization. These limitations can lead to skewed perceptions of AI’s real-world use and value[52].

Even within an active community, benchmarks are unable to capture the full array of AI system functionality and performance. Benchmarking can lag behind new model capabilities, especially for complex agentic tasks or qualitative aspects like creativity and reasoning. Benchmarks can be prone to task contamination or data leakage resulting in erroneously high performance [51, 60]. The intense interest in generative AI model capabilities has driven the use of tests that were designed for other purposes (e.g., testing models on college admission tests or professional certifications) – or poorly specify human tasks, potentially distorting perceptions of progress[47, 67, 3]. Benchmarks are designed to mathematically represent complex human and societal phenomena, which can contribute to a fallacy of objectivity [83]. Benchmarking metrics focus on system accuracy or policy violations, which are challenging to apply to second order questions.

Arguably the biggest weakness of benchmarking is its inability to account for the inter-dependencies between humans and AI, such as how people leverage AI or interpret and act upon AI-generated output in the real world, and what it means at a societal level. Even the most comprehensive benchmark suites remain abstractions that offer only partial glimpses into real world effects[71]. The need for

scale has led to reliance on static benchmark datasets and highly constrained tasks which are a poor match for deployment environments where contextual factors and user perceptions can dramatically alter outcomes [107, 27, 33, 85].

3 Context is Everything: Crafting a Real World AI Evaluation Ecosystem

The ability to make claims about the real world requires authentic and extensive contextual detail. Contextual awareness - knowledge about what matters in a given deployment setting - can improve AI's fit within societal contexts and foster measurement validity. Contextual information can fulfill two requirements for real world AI evaluation that benchmarking struggles to address. First, non-ML actors use this information to translate and make sense of evaluation results for their own activities and decision making. Second, practitioners on the AI stack can gain complementary evidence of how the technology they build is actually being used in deployment.

Currently, sensing and leveraging contextual information from the real world is impeded by processes in the AI stack. While ML models can be derived from trillions of data points, the development process flattens contextual detail. Recent approaches to align model outcomes to predefined and prescriptive values [5] reduce societal and contextual detail instead of eliciting and analyzing it. Many organizations also lack the skills and methods to interpret and translate contextual material from the real world (such as user reviews, information from redress and recourse, other stakeholder feedback) into AI product workflows[90]. Combined, these practices can bake in brittle performance once AI systems are deployed [83, 16, 55, 21, 56, 66].

This section describes methods for how to specify context for real world evaluation and to collect and generate contextually-informed data. Methods for analyzing contextual information will be the focus of future directions.

Establishing Contextual Awareness
The field of value-sensitive design (VSD) and its tripartite methodology (conceptual, empirical, technical) provides a foundational framework to operationalize contextual awareness for real world AI evaluation. Table 2 summarizes how practices for specifying contextual scope and collecting real world data fit into the VSD framework.

Contextual Approach	Key Integration Practices	Outcome	VSD Method
Context Specification	Initiate Theory of Change Systematize Real World Concepts Stakeholder Engagement	Contextually informed requirements for data collection and generation activities.	Conceptual
Data Collection and Generation	Field Testing Red Teaming	Data about regular and adversarial use of AI systems.	Empirical Technical

Table 2: Overview of context-aware AI evaluation approaches and their interdisciplinary roles.

By docking into the VSD framework, real world AI evaluation methods can produce continuous feedback loops–where context specification activities inform red teaming (to identify real-world failures) and field testing (to determine extent of failures in regular use). Since red teaming and field testing enable investigation of "the technology, the people who use it, and the social systems that configure, use, or are otherwise affected by the technology"[38] it satisfies both the empirical and technical VSD methods. VSD processes can also assist in translating evaluation outcomes into technical/policy adjustments.

3.1 Context Specification Activities

The activities described below define the real world challenge problem, the context in which it exists, and other relevant detail. Gathering this information is the first step in facilitating contextual awareness and requires input from a broad set of stakeholders to ensure measurement validity.

3.1.1 Theory of Change

Real world AI evaluation activities are initiated by defining a theory of change. Key stakeholders and evaluators collaboratively identify challenge problems, specify desired goals over the current state and determine evaluation inputs, activities, outputs, and outcomes. Stakeholders also assist evaluators in identifying counterfactuals to estimate what might happen without the evaluation effort.

3.1.2 Systematization of Real World Concepts

Real world concepts that underlie the development of an AI model's objective function and other variables drive system functionality, optimization and performance. The validity of an AI model can hinge on how well these real world concepts are systematized and operationalized [26], which requires technical, neutral, collectively informed and unambiguous descriptions. Models that do not demonstrate validity cannot maintain performance well across contexts. Systematized descriptions can be used to:

- instruct AI models to properly recognize a given phenomenon and act accordingly in context,
- optimize development of prompts for user engagement with AI systems and to ensure model outcome meets preferences and requirements,
- enhance content markup and moderation for complex and ambiguous phenomena (e.g., obscenity, abusive or hateful content).

Currently, ML practitioners demonstrate difficulty with systematization and operationalization, and it is challenging to bridge the communication divide between computational and other disciplines and translate real world concepts along product lifecycles [36, 91, 34, 66, 92] .

3.1.3 Stakeholder Feedback and Adaptive Governance

AI evaluators are increasingly exploring methods that better reflect deployment conditions and integrate members of the public directly into the measurement process. Meaningful stakeholder engagement methods are a common component of adaptive AI governance frameworks [59, 28, 11] and can bolster public accountability, democratic governance, and transparency efforts such as recourse and redress. Engagement is conducted throughout the entire AI project lifecycle and can effectively inform evaluation activities. Engagement activities use a variety of qualitative methods to capture a range of perspectives and experiences from stakeholders external to the AI development organization. Stakeholder engagement activities can be built into evaluation paradigms to facilitate contextual awareness [57, 8, 68, 48] by:

- revealing potential negative impacts prior to AI development and deployment and shed light on unanticipated AI uses and positive outcomes,
- surfacing emergent risks or gradual declines in real world system performance
- informing mitigation of AI harms before they become entrenched, [62, 85, 4]
- surfacing assumptions and limitations about AI technology.

The Alan Turing Institute's AI Sustainability in Practice workbook [59] lays out a stakeholder engagement process which begins with a determination of the groups most likely to be negatively impacted by AI systems. The level of subsequent stakeholder involvement–ranging from inform or consult to partner or empower–is proportionate to the scope of a project's potential risks and impacts [59]. Participatory co-creation is another engagement method that moves beyond traditional consultation to enable and empower stakeholders in more active roles across the AI design, development, deployment processes. Stakeholders work closely with AI designers from the initial context specification phase, iterate on the design and user interface, support the creation of governance structures, and inform system monitoring [8].

3.2 Collecting and Generating Contextually-Informed Data

Once the contextual unit of interest has been defined, data collection activities can be designed and executed. Two methods for collecting and generating contextually informed data – field testing and red teaming– are described below. While benchmarking relies almost entirely on curated and labeled

datasets, red teaming and field testing can be used to design and collect response data from different types of audiences as they interact with AI systems in the real world.

3.2.1 Field Testing

Field methods and experiments have been used by social scientists for decades to gain insights into human and social behavior by bridging laboratory settings and the real world. Methods similar to field testing[1] are regularly used in technology settings but its adapted use in AI evaluation is relatively nascent, with recent work in the field of AI risk assessment [84, 79]. Designed to elicit and capture detailed information about what happens under regular use, field testing is conducted through empirical observation of individuals as they interact with AI technologies under semi-controlled conditions across multiple sessions. While the focus of benchmarking is the AI model or system, AI field testing can focus on the "contextual unit" – or the complex and adaptive behavior that naturally occurs as people leverage AI technology in setting. Field testing can be used to explore how humans use and adapt to AI technology, investigate feedback loops between humans and technology, [27, 98, 41], and uncover emergent or “long-tail” scenarios that single-turn, lab-based benchmarks might miss.

In a simulated sandbox and reporting environment[2], hundreds or even thousands of human subjects interact live with AI systems and provide feedback about their experiences and subsequent actions. Resulting dialogues from test interactions can be annotated to determine whether various phenomena materialized. This descriptive reporting approach transforms evaluation paradigms beyond whether or not a system generated "the right answer" or asking people to judge AI output or train AI systems. Instead, field testing enables the collection of real world evidence about what materializes when certain AI features are deployed to the broader public. Since field tests are conducted in a controlled and protected environment, evaluators can safely configure pre-deployment testing suites and responsibly explore a wide variety of factors. When using field testing to measure accuracy of system responses, task contamination and data leakage are less likely than in benchmarking due to the difficulty of anticipating the heterogeneous prompts of thousands of testers.

Field testing requires:

- Multi-session experiments to observe how subjects adapt to AI technology over repeated usage (days or weeks).
- Experimental randomization and blinding to minimize biases in user interactions or system responses.
- Observation and analysis of subject responses and behaviors alongside isolated system outputs such as user surveys, logs, and performance metrics.
- Test scenarios for subject interactions with AI systems that balance naturalistic conditions and subject safety [79].
- Human subject research protocols.
- Descriptive approaches for marking up interactive output [79].

3.2.2 Red Teaming

The rise in generative AI use and its associated impacts has contributed to increased interest in AI red teaming as a complement to conventional evaluation paradigms[108]. Unlike static benchmarks, red teaming can simulate real-world usage to

- uncover failures, trends and patterns that emerge in complex or adversarial settings,
- highlight misuse and weaknesses in system behavior and robustness,
- determine boundary conditions to inform go/no-go decisions about deploying AI, and
- verify the effectiveness of existing mitigation strategies, safety measures and frameworks.

Red teaming is often conducted via "challenges", where individual testers use simulated attacks to identify vulnerabilities and evaluate the safety and security of AI systems. Red teamers may use

[1] such as A/B tests

[2] Can also be referred to as a large-scale human testbed

creative multi-turn prompting, role-playing, and other techniques to probe the model's responses and surface undesirable model outputs, such as data leakage [1], jailbreaking [2], and information based harms[3] Red teaming challenges can surface detailed information about how harmful outcomes occur, who they affect, how they circulate in social contexts or are repurposed by malicious actors, and how system vulnerabilities evolve over time [98, 22]. Red teaming is especially valuable in high-stakes domains like education, healthcare, and employment, where harms may be severe or emerge gradually, or disproportionately impact marginalized groups.

Red teaming challenges require detailed instructions, rules of engagement, and a framework, policy, or set of rules for identifying violative outcomes. Various tasks along the AI pipeline may require individuals to engage with harmful test scenarios or to be exposed to toxic and violent content, and red teaming is no different. To protect red teamer safety, challenges require appropriate psychological safety mechanisms to be put in place prior to participant enrollment.

Red teaming requires diverse backgrounds and domain expertise to cover the broad range of potential harms posed by AI systems. For example, multi-lingual expertise is required to test AI systems for linguistic and dialectal biases and gaps in language coverage. Challenges can go beyond simple Q&A tasks to test models on summarization and translation tasks and sentiment analysis. Red Teaming challenges may entail:

- **Expert Red Teaming:** Highly skilled professionals with expertise in adversarial misuse or exploits, or in the underlying subject matter, simulate sophisticated attacks to identify deep-seated vulnerabilities.
- **Public Red Teaming:** Members of the general public interact with AI systems under controlled or "challenge" conditions to complement expert red teaming and expand the tested risk surface. Public participants do not require expertise in adversarial testing but instead seek to surface real-world failures or "off-label" uses that expert red teamers may not anticipate or consider, such as how AI systems may fail across cultural or linguistic contexts.
- **Automated Red Teaming:** The automated generation of adversarial prompts or test cases at scale to uncover issues such as data leakage or content policy circumvention. Evaluators can automate parts of the red teaming process to expand test coverage and reveal systemic model weaknesses.

Challenge designers can combine public and expert-based red teaming exercises into hybrid challenges and leverage principles of collective intelligence, where testers can coordinate with – or learn from – each other's discoveries.[4] Collaborative and asynchronous exercises can encourage knowledge-sharing and expedite the discovery of edge cases. Manual and automated techniques can also be combined to balance the strengths and limitations of both approaches[69]. Red teaming can be used alongside field testing to determine whether adversarial vulnerabilities may manifest in regular use, or if new ones arise from repeated user queries

Red Teaming Attack Strategies In addition to the list of red teaming attack strategies found in Appendix A, red teamers can systematically employ data poisoning, indirect prompt injection, or multi-turn "scenario chaining" to force AI systems into unforeseeable states and capture vulnerabilities that may only appear after multiple interactions or under disguised prompts. Periodic red teaming "rounds" can be used to track whether system updates inadvertently open up new exploits or degrade previously solved safeguards.

Selected Red Teaming Limitations As AI systems evolve, red teaming efforts can adapt through interdisciplinary development of new attack vectors and multi-turn or multi-modal tasks. [46, 98]. A list of recommendations that challenge designers can use to address selected red teaming limitations is provided below:

[1] Revealing sensitive information from AI system training data.

[2] Circumventing safety measures and generating restricted, privileged, dangerous, copyrighted and/or otherwise unauthorized material.

[3] Obscene, degrading, abusive, and radicalizing material; content that may not distinguish fact from fiction; content that may amplify, reify or exacerbate biases against different sub-groups or lead to disparities between sub-groups; false content that may mislead or deceive users (aka hallucinations).

[4] Small groups of experts can collectively overcome a learning curve faster than individuals, allowing them to identify more subtle or complex vulnerabilities in a shorter time frame.[22, 95]

- **Scoping** Address scoping limitations by including multi-turn conversations, multiple languages and dialects, and multi-modal tasks.
- **Tester Biases** Address participant bias and representation issues by expanding the red teaming recruitment process beyond traditional settings, broadening dataset requirements, surveying red teamer perceptions of harm, and introducing positionality statements.
- **Automation** Collaboratively develop criteria for automated generation of high-quality, diverse test cases while preserving the nuanced understanding of human red teamers.
- **Resource Constraints** Balance the cost and efficiency of manual red teaming with scalable but limited automated approaches to ensure engagement from smaller organizations or research groups.
- **Transparency and Information Sharing** Establish guidelines for the responsible, open and transparent sharing of red teaming findings that take ethical implications and potential misuse into account.
- **Evaluating Effectiveness** Build off of information security red teaming metrics[1] to collaboratively define criteria for desirable and undesirable system behavior and advance evaluation metrics and methods to track progress over time.

4 Summary and Recommendations

Policy makers, organizational decision makers and members of the public each require different types of information about AI so they can make informed decisions about whether and how to develop, deploy or use it in their own contexts. A real world AI evaluation ecosystem to support these audiences will have to contend with many trade-offs to gather information beyond the AI stack and within context.

While benchmarking is too limited on its own to investigate second order effects, other types of evaluation that provide more fidelity are disconnected from the necessary system measurements central to AI benchmarking. "Contextual work" is commonly viewed as slow and resource-intensive compared to benchmarking, since it requires different processes, actors, skills and disciplines. For example, fielding qualitative research surveys and conducting ethnographies are both more expensive and time-consuming than using "found data". Activities surrounding problem specification are also consistently overlooked due to a perception that they take too long and don't provide enough benefit.

Both red teaming and field testing require infrastructure that can host people and technology in deployed scenarios while meeting human subject research requirements. All evaluation methods will require transparency, reproducibility, and scientific integrity. Even when built on feedback from thousands of people, evaluation outcomes do not automatically ladder up to societal insights such as impacts to democracy, the workforce and the economy, education, and culture[45, 97, 15, 2, 65].

With no existing infrastructure or community dedicated to evaluating AI's second order effects, other procedural models could be used as exemplars. A new ecosystem could be supported through the creation of testing hubs that include expertise from academia, industry, and civil society to develop rigorous science-backed evaluation methodologies and frameworks. Ecosystem inputs could be sourced from organizations that bring their questions to bear. Members of the public could support specification of contextual inputs and enroll in red teaming and field testing activities. Organizations that have relevant evaluation expertise and methods can provide their services as independent testers to enhance credibility and ensure objectivity in the evaluation process. The academic research community can support the development of formalized metrics and methodologies. Over time, the ecosystem can determine which evaluation activities produce value and should be automated (and semi-automated) to enhance scalability and adoption.

Outputs from ecosystem activities will center on answering second order effects and fostering a more dynamic and adaptive real world AI evaluation community. Anticipated insights will include deeper understanding of how AI technologies function outside tightly controlled lab settings, how users might abuse or misunderstand AI functionality and outputs, and how AI's role in society influences systemic trends.

[1] such as incident detection rate, time to detect incident, and mean time to recovery

References

[1] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. A collaborative, human-centred taxonomy of ai, algorithmic, and automation harms. *arXiv preprint arXiv:2407.01294*, 2024.

[2] Daron Acemoglu. The simple macroeconomics of ai. Working Paper 32487, National Bureau of Economic Research, May 2024. URL http://www.nber.org/papers/w32487.

[3] Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. Which humans?

[4] U. Aïvodji et al. Fairwashing: the risk of rationalization. In *Proceedings of the International Conference on Machine Learning*, pages 161–170, 2019. doi: 10.48550/arXiv.1901.09749. URL https://doi.org/10.48550/arXiv.1901.09749.

[5] Y. Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[6] Loukas Balafoutas, Jeremy Celse, Alexandros Karakostas, and Nicholas Umashev. Incentives and the replication crisis in social sciences: A critical review of open science practices. *Journal of Behavioral and Experimental Economics*, 114:102327, 2025. ISSN 2214-8043. doi: https://doi.org/10.1016/j.socec.2024.102327. URL https://www.sciencedirect.com/science/article/pii/S2214804324001642.

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[8] Aleks Berditchevskaia, Eirini Malliaraki, and Kathy Peach. Participatory ai for humanitarian innovation: A briefing paper. *Nesta*, 2021. URL https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf.

[9] M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *The American Economic Review*, 94 (4):991–1013, 2004. URL https://www.jstor.org/stable/3592802.

[10] Samuel Bilson, Maurice Cox, Anna Pustogvar, and Andrew Thompson. A metrological framework for uncertainty evaluation in machine learning classification models, May 2025. URL http://arxiv.org/abs/2504.03359. arXiv:2504.03359 [cs].

[11] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555290. URL https://doi.org/10.1145/3551624.3555290.

[12] B. Blaire et al. Legal red teaming: A systematic approach to assessing legal risk of generative ai models. https://www.dlapiper.com/-/media/project/dlapiper-tenant/dlapiper/pdf/dla-piper---white-paper---ai-legal-red-teaming.pdf, 2024.

[13] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. Atlas of ai risks: Enhancing public understanding of ai risks. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 12, pages 33–43, 2024.

[14] R. Bommasani, P. Liang, and T. Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: 10.1111/nyas.15007. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/nyas.15007.

[15] Kathryn Bonney, Cory Breaux, Catherine Buffington, Emin Dinlersoz, Lucia Foster, Nathan Goldschlag, John Haltiwanger, Zachary Kroff, and Keith Savage. The impact of ai on the workforce: Tasks versus jobs? *Economics Letters*, 244:111971, 2024. ISSN 0165-1765. doi: https://doi.org/10.1016/j.econlet.2024.111971. URL https://www.sciencedirect.com/science/article/pii/S0165176524004555.

[16] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. Overcoming failures of imagination in ai infused system development and deployment, 2020. URL https://arxiv.org/abs/2011.13416.

[17] Colin Camerer, Anna Dreber, Felix Holzmeister, Teck Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, and Hang Wu. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2, 09 2018. doi: 10.1038/s41562-018-0399-z.

[18] N. Carlini et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[19] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models, 2023. URL https://arxiv.org/abs/2202.07646.

[20] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model, 2024. URL https://arxiv.org/abs/2403.06634.

[21] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

[22] D. Centola. The network science of collective intelligence. *Trends in Cognitive Sciences*, 26 (11):923–941, 2022. URL https://pubmed.ncbi.nlm.nih.gov/36180361/.

[23] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[24] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

[25] Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Wei, Afaf Taik, Jackie Chi Kit Cheung, and Golnoosh Farnadi. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15694–15710, 2024.

[26] Alexandra Chouldechova, Chad Atalla, Solon Barocas, A. Feder Cooper, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. A shared standard for valid measurement of generative ai systems' capabilities, risks, and impacts, 2024. URL https://arxiv.org/abs/2412.01934.

[27] A. D'Amour et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. doi: 10.48550/arXiv.2011.03395. URL https://doi.org/10.48550/arXiv.2011.03395.

[28] Cedric E. Dawkins. The principle of good faith: Toward substantive stakeholder engagement. *J Bus Ethics*, 121:283–295, 2014. doi: 10.1007/s10551-013-1697-z.

[29] R. Dobbe, T. K. Gilbert, and Y. Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021. doi: 10.48550/arXiv.2106.11022. URL https://arxiv.org/abs/2106.11022.

[30] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results, 2019. URL https://arxiv.org/abs/1909.03004.

[31] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–12, 2024.

[32] M. Duan et al. Do membership inference attacks work on large language models? arXiv preprint arXiv:2402.07841, 2024. Feb 12.

[33] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyên Hoang, Rafael Pinot, Sébastien Rouault, and John Stephan. On the impossible safety of large ai models, 2023. URL https://arxiv.org/abs/2209.15259.

[34] Moss Emanuel and Metcalf Jacob. *Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies*. Data Society, September 2020. URL https://datasociety.net/wp-content/uploads/2020/09/Ethics-Owners_20200923-DataSociety.pdf.

[35] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation, 2025. URL https://arxiv.org/abs/2502.06559.

[36] Nel Escher, Jeffrey Bilik, Nikola Banovic, and Ben Green. Code-ifying the law: How disciplinary divides afflict the development of legal software. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2), November 2024. doi: 10.1145/3686937. URL https://doi.org/10.1145/3686937.

[37] Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. Large ai models are cultural and social technologies. *Science*, 387(6739):1153–1156, 2025. doi: 10.1126/science.adt9819. URL https://www.science.org/doi/abs/10.1126/science.adt9819.

[38] Batya Friedman, Peter Kahn, and Alan Borning. *UW CSE Technical Report*.

[39] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, October 2016. URL http://arxiv.org/abs/1506.02142. arXiv:1506.02142 [stat].

[40] Gallup and Telescope Foundation. Americans use ai in everyday products without realizing it. URL https://www.telescopegp.com/insights/americans-use-ai-in-everyday-products-without-realizing-it.

[41] L. Gao et al. A framework for few-shot language model evaluation. https://github.com/EleutherAI/lm-evaluation-harness, 2021.

[42] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[43] Nicole Gillespie, Steven Lockey, Tabi Ward, Alexandria Macdade, and Gerard Hassed. Trust, attitudes and use of artificial intelligence: A global study 2025, 2025. URL https://figshare.unimelb.edu.au/articles/report/Trust_attitudes_and_use_of_artificial_intelligence_A_global_study_2025/28822919.

[44] K. Greshake et al. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90, 2023. URL https://arxiv.org/abs/2302.12173.

[45] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL https://arxiv.org/abs/2503.04761.

[46] M. Hoffmann and H. Frase. Adding structure to ai harm: An introduction to cset's ai harm framework. https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/, 2023.

[47] Anna A. Ivanova. Toward best research practices in ai psychology, 2024. URL https://arxiv.org/abs/2312.01276.

[48] Sikke R. Jansma, Anne M. Dijkstra, and Menno D.T. de Jong. Co-creation in support of responsible research and innovation: an analysis of three stakeholder workshops on nanotechnology for health. *Journal of Responsible Innovation*, 9(1):28–48, 2021. doi: 10.1080/23299460.2021.1994195.

[49] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[50] S. Kapoor and A. Narayanan. Openai's policies hinder reproducible research on language models. https://www.aisnakeoil.com/p/openais-policies-hinder-reproducible, 2023.

[51] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science, 2022. URL https://arxiv.org/abs/2207.07048.

[52] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

[53] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

[54] Max Korbmacher, Flávio Azevedo, Charlotte Pennington, Helena Hartmann, Madeleine Pownall, Kathleen Schmidt, Mahmoud Elsherif, Nate Breznau, Olly Robertson, Tamara Kalandadze, Iris Shijun Yu, Bradley Baker, Aoife O'Mahony, Jørgen Olsnes, John Shaw, Biljana Gjoneska, Yuki Yamada, Jan Röer, Jennifer Murphy, and Thomas Evans. The replication crisis has led to positive structural, procedural, and community changes, 05 2023.

[55] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback, 2023. URL https://arxiv.org/abs/2310.13595.

[56] Susan Landau, James X. Dempsey, Ece Kamar, Steven M. Bellovin, and Robert Pool. Challenging the machine: Contestability in government ai systems, 2024. URL https://arxiv.org/abs/2406.10430.

[57] Calum F. Leask, Marlene Sandlund, Dawn A. Skelton, Teatske M. Altenburg, Greet Cardon, et al. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. *Res Involv Engagem*, 5(2): 1153–1156, 2019. doi: 10.1186/s40900-018-0136-9.

[58] B. Lenaerts-Bergmans. Data poisoning: The exploitation of generative ai. https://www.crowdstrike.com/cybersecurity-101/cyberattacks/data-poisoning/, 2024.

[59] David Leslie, Cami Rincon, Morgan Briggs, et al. Ai sustainability in practice part one: Foundations for sustainable ai projects. https://aiethics.turing.ac.uk/modules/sustainability-1/, 2024.

[60] C. Li and J. Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18471–18480, 2024. doi: 10.48550/arXiv.2312.16337. URL https://doi.org/10.48550/arXiv.2312.16337.

[61] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. LLM defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024.

[62] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

[63] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

[64] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/757b505cfd34c64c85ca5b5690ee5293-Paper-round2.pdf.

[65] Brent Lutes, Joshua Gans, Shane Greenstein, Adam B. Jaffe, Abhishek Nagaraj, Imke Reimers, Michael D. Smith, Rahul Telang, Catherine E. Tucker, and Joel Waldfogel. Identifying the economic implications of artificial intelligence for copyright policy (february 12, 2025). first published by the u.s. copyright office. *First published by the U.S. Copyright Office, Available at SSRN: https://ssrn.com/abstract=5143605 or http://dx.doi.org/10.2139/ssrn.5143605*, 2025.

[66] Michael A. Madaio, Jingya Chen, Hanna Wallach, and Jennifer Wortman Vaughan. Tinker, tailor, configure, customize: The articulation work of contextualizing an ai fairness checklist. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), April 2024. doi: 10.1145/3653705. URL https://doi.org/10.1145/3653705.

[67] E. Martínez. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 30:1–24, 2024. doi: 10.1177/20539517241290220. URL https://doi.org/10.1177/20539517241290220.

[68] Cristian Matti and Gabriel Rissola. Co-creation for policy: Participatory methodologies to structure multi-stakeholder policymaking processes. https://www.eit.europa.eu/sites/default/files/jrc128771_01.*pdf*, 2022.

[69] A. Mislove. Red-teaming large language models to identify novel ai risks. https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/, 2023.

[70] Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.0/.

[71] A. Nelson. Facct'23 keynote: "thick alignment". https://www.youtube.com/watch?v=Sq_XwqVTqvQ, 2023.

[72] International Association of Privacy Professionals. Global law and policy tracker. URL https://iapp.org/resources/article/global-ai-legislation-tracker/.

[73] OECD Artificial Intelligence Papers. Governing with artificial intelligence, June 2024. URL https://www.oecd.org/en/publications/governing-with-artificial-intelligence_26324bc2-en.html.

[74] Luona Lin Parker and Kim. U.s. workers are more worried than hopeful about future ai use in the workplace, February 2025. URL https://www.pewresearch.org/social-trends/2025/02/25/u-s-workers-are-more-worried-than-hopeful-about-future-ai-use-in-the-workplace/.

[75] Brian Kennedy Jeffrey Gottfried Monica Anderson Pasquini, Colleen McClain and Giancarlo. How the u.s. public and ai experts view artificial intelligence, April 2025. URL https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/.

[76] Inioluwa Deborah Raji and Roel Dobbe. Concrete problems in ai safety, revisited, 2023. URL https://arxiv.org/abs/2401.10899.

[77] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

[78] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

[79] Razvan Amironesei Craig Greenberg Jon Fiscus Patrick Hall Anya Jones Shomik Jain Afzal Godil Kristen Greene Ted Jensen Reva Schwartz, Gabriella Waters and Noah Schulman. The assessing risks and impacts of ai (aria) program evaluation design document. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, 2024.

[80] Anna Rogers, Tim Baldwin, and Kobi Leins. Just what do you think you're doing, dave?' a checklist for responsible data use in nlp, 2021. URL https://arxiv.org/abs/2109.06598.

[81] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

[82] Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024. URL https://arxiv.org/abs/2402.16822.

[83] R. Schwartz et al. Towards a standard for identifying and managing bias in artificial intelligence. Technical Report NIST SP 1270, National Institute of Standards and Technology, Gaithersburg, MD, 2022. URL https://doi.org/10.6028/NIST.SP.1270.

[84] Reva Schwartz et al. The nist assessing risks and impacts of ai (aria) pilot evaluation plan. Technical report, National Institute of Standards and Technology, Gaithersburg, MD, 2024. https://ai-challenges.nist.gov/aria/docs/evaluation_plan.pdf.

[85] A. D. Selbst et al. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

[86] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, 2023.

[87] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. Taskbench: Benchmarking large language models for task automation. *Advances in Neural Information Processing Systems*, 37:4540–4574, 2024.

[88] I. Shumailov et al. Sponge examples: Energy-latency attacks on neural networks. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 212–231, 2021. URL https://arxiv.org/abs/2006.03463.

[89] Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion, 2025. URL https://arxiv.org/abs/2504.20879.

[90] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394772. doi: 10.1145/3551624.3555285. URL https://doi.org/10.1145/3551624.3555285.

[91] Stephen C. Slota, Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. Many hands make many fingers to point: Challenges in creating accountable ai. *AI and Society*, 38(4):1287–1299, 2023. doi: 10.1007/s00146-021-01302-0.

[92] Jessie J. Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. Real ml: Recognizing, exploring, and articulating limitations of machine learning research. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, page 587–597. ACM, June 2022. doi: 10.1145/3531146.3533122. URL http://dx.doi.org/10.1145/3531146.3533122.

[93] A. Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

[94] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models, 2024. URL https://arxiv.org/abs/2310.07298.

[95] C. Stadler, T. Rajwani, and F. Karaba. Solutions to the exploration/exploitation dilemma: networks as a new level of analysis. *International Journal of Management Reviews*, 16(2):172–193, 2013. doi: 10.1111/ijmr.12015. URL https://doi.org/10.1111/ijmr.12015.

[96] Ilan Strauss, Isobel Moure, Tim O'Reilly, and Sruly Rosenblat. *The State of AI Governance Research: AI Safety and Reliability in Real World Commercial Deployment*. AI Disclosures Project, Social Science Research Council, April 2025. doi: 10.35650/aidp.4112.d.2025. URL http://dx.doi.org/10.35650/AIDP.4112.d.2025.

[97] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, and Deep Ganguli. Clio: Privacy-preserving insights into real-world ai use, 2024. URL https://arxiv.org/abs/2412.13678.

[98] M. Theofanos, Y. Choong, and T. Jensen. Ai use taxonomy: A human-centered approach. Technical Report NIST AI 200-1, National Institute of Standards and Technology, Gaithersburg, MD, 2024. URL https://doi.org/10.6028/NIST.AI.200-1.

[99] R. Thomas and D. Uminsky. Reliance on metrics is a fundamental challenge for ai. https://arxiv.org/pdf/2002.08512, 2020.

[100] Andrew Thompson. Analytical results for uncertainty propagation through trained machine learning regression models, May 2024. URL http://arxiv.org/abs/2404.11224. arXiv:2404.11224 [cs].

[101] E. L. Trist and K. W. Bamforth. Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, 4 (1):3–38, February 1951. ISSN 0018-7267, 1741-282X. doi: 10.1177/001872675100400101. URL https://journals.sagepub.com/doi/10.1177/001872675100400101.

[102] UK Department for Science, Innovation and Technology. Prime minister sets out blueprint to turbocharge ai. https://https://www.gov.uk/government/news/prime-minister-sets-out-blueprint-to-turbocharge-ai, 2025.

[103] Matias Valdenegro-Toro and Daniel Saromo. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement, April 2022. URL http://arxiv.org/abs/2204.09308. arXiv:2204.09308 [cs].

[104] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative ai systems is a social science measurement challenge, 2025. URL https://arxiv.org/abs/2502.00561.

[105] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023.

[106] Cheng Wang. Calibration in Deep Learning: A Survey of the State-of-the-Art, May 2024. URL http://arxiv.org/abs/2308.01222. arXiv:2308.01222 [cs].

[107] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023. URL https://arxiv.org/abs/2310.11986.

[108] Alice Qian Zhang, Ryland Shaw, Jacy Reese Anthis, Ashlee Milton, Emily Tseng, Jina Suh, Lama Ahmad, Ram Shankar Siva Kumar, Julian Posada, Benjamin Shestakofsky, Sarah T. Roberts, and Mary L. Gray. The human factor in ai red teaming: Perspectives from social and collaborative computing, 2024. URL https://arxiv.org/abs/2407.07786.

[109] Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523, 2023.

[110] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Appendix

A Selected Red Teaming Attack Strategies

- Complex or leading prompts can expose common AI system vulnerabilities such as confabulation, logically inconsistent responses, faulty reasoning, flawed decision-making, incorrect numeric responses, erroneous code generation, and fabricated citations.
- Counterfactual prompting and the use of repeated requests while varying demographic personas can uncover harmful biases [9].
- Autocompletion, fill-in-the-blank requests and prompts designed as "honest" requests can be used to evaluate system guardrails and force AI systems to produce harmful completions.
- Membership inference attacks, and probes of training data memorization can be used to expose sensitive or private information [18–20, 32, 94].
- Prompting for sensitive personal or location-based details can be used to evaluate data handling and privacy safeguards.
- Combining jailbreaking attacks with counterfactual prompts in multiple languages and dialects can be used to force culturally and linguistically biased output.
- Data poisoning, indirect prompt injection, misleading training inputs [58], and embedding harmful prompts subtly within benign content [44] can be used to evaluate system integrity and resistance to manipulation.
- Availability or "sponge" attacks use excessively large numbers of queries to stress test AI systems for performance stability and resource resilience [88].
- Chaos testing and random attacks expose systems to excessively large numbers of random prompts to elicit failures or jailbreaks (these prompts can be AI generated).
- Adversarial examples and membership inference attacks are used to probe security vulnerabilities [18–20, 32, 94].
- Prompts for copyrighted or proprietary content can be used to surface intellectual property risks [12].
- Prompts for obscene or abusive content can be used to evaluate the efficacy of content moderation.