Toward an evaluation science for generative AI systems

Laura Weidinger[*]
Google DeepMind

Inioluwa Deborah Raji[*]
University of California, Berkeley

Hanna Wallach
Microsoft Research

Margaret Mitchell
Hugging Face

Angelina Wang
Stanford University

Olawale Salaudeen
Massachusetts Institute of Technology

Rishi Bommasani
Stanford University

Deep Ganguli
Anthropic

Sanmi Koyejo
Stanford University

William Isaac
Google DeepMind

Abstract

There is an increasing imperative to anticipate and understand the performance and safety of generative AI systems in real-world deployment contexts. However, the current evaluation ecosystem is insufficient: commonly used static benchmarks face validity challenges, and ad hoc case-by-case approaches rarely scale. In this piece, we advocate for maturing an evaluation *science* for generative AI systems. While generative AI creates unique challenges for system safety engineering and measurement science, the field can draw valuable insights from the development of safety evaluation practices in other fields including transportation, aerospace, and pharmaceutical engineering. In particular, we present three key lessons: evaluation metrics must be applicable to real-world performance, metrics must be iteratively refined, and evaluation institutions and norms must be established. Applying these insights, we outline a concrete path toward a more rigorous approach for evaluating generative AI systems.

---

[*]= contributed equally to this work

1 Introduction

The widespread deployment of generative AI systems in medicine (Boyd 2023), law (e.g., Lexis+AI[1]), education (Singer 2023b), information technology (e.g., Microsoft's Copilot [2], Reid (2024), and many social settings (e.g. Replika[3], character.AI) has led to a collective realization: the performance and safety of generative AI systems in real-world deployment contexts are very often poorly anticipated and understood (Roose 2024a; Mulligan; Wiggers 2024). The tendency of these systems to generate inaccurate statements has already led to the spreading of medical and other misinformation (Omiye et al. 2023; Archer and Elliott 2025); incorrect legal references (Magesh et al. 2024); failures as educational support tools (Singer 2023a); and widespread confusions in search engine use (Heaven 2022; Murphy Kelly 2023). Beyond factual discrepancies, AI-enabled chatbots have also been described as interacting inappropriately with users (Roose 2023), exposing security vulnerabilities (Nicholson 2023), and fostering unhealthy emotional reliance (Roose 2024b; Dzieza 2024; Verma 2023).

The historical focus on benchmarks and leaderboards has been effective at encouraging the AI research community to pursue shared directions; however, as AI products become widely integrated into our everyday lives, it is increasingly clear that static benchmarks are not well suited to improving our understanding of the *real-world* performance and safety of deployed generative AI systems (Bowman and Dahl 2021; Goldfarb-Tarrant et al. 2021a; Liao et al. 2021; Raji 2021; de Vries et al. 2020). Despite this mismatch, static benchmarks are still commonly used to inform real-world decisions about generative AI systems that stretch far beyond the research landscape – such as in deployment criteria and marketing materials for new model or system releases (Grattafiori et al. 2024; Anthropic 2024b; OpenAI 2024; Gemini et al. 2024), third party critiques (Mirzadeh et al. 2024; Zhang et al. 2024), procurement guidelines (Johnson et al. 2025), and in public policy discourse (European Commission 2024; National Institute of Standards and Technology 2021). Although there is an emerging interest in more interactive (Chiang et al. 2024), dynamic (Kiela et al. 2021), and behavioral (Ribeiro et al. 2020) approaches to evaluation, many of the existing alternatives to benchmarks, such as red teaming exercises and case-by-case audits, still fall woefully short of enabling systematic assessments and accountability (Birhane et al. 2024; Friedler et al. 2023).

For AI evaluation to mature into a proper "science," it must meet certain criteria. Sciences are marked by having theories about their targets of study, which can be expressed as testable hypotheses. Measurement instruments to test these hypotheses must provide experimental consistency (i.e., reliability, internal validity) and generalizability (i.e., external validity). Finally, sciences are marked by iteration: Over time, measurement approaches and instruments are refined and new insights are uncovered. Collectively, these properties of sciences contrast sharply with the practice of rapidly developing static benchmarks for evaluating generative AI systems, while anticipating that within a few months such benchmarks will become much less useful or obsolete.

As generative AI exits an era of research and enters a period of widespread use (Hu 2023; Reid 2024), the field risks exacerbating an ongoing public crisis of confidence in AI technology (Faverio and Tyson 2023) if we do not develop a more mature evaluation science for generative AI systems. From the history of other fields, we can get a sense of why: Collectively, leaderboards, benchmarks, and audits do not amount to the robust and meaningful evaluation ecosystem we need to properly assess the suitability of these products in widespread use. In particular, they cannot give assurances about AI system performance in different domains or for different user groups[4]. In this piece, we advocate for the maturation of such an evaluation science. By drawing on insights from systems safety engineering and measurement science in other fields, while acknowledging the unique challenges inherent to generative AI, we identify three important properties of any evaluation science that the AI community will need to focus on to meaningfully advance progress: a focus on real-world applicability, iterative measurement refinement, and adequate institutional

---

[1]https://www.lexisnexis.com/en-us/products/lexis-plus-ai.page

[2]copilot.microsoft.com

[3]replika.com

[4]We thank our anonymous reviewer for pointing out that one reason for this is that benchmarks are often not robust to data or domain shifts, i.e. benchmarks test AI system outputs in certain contexts but this may not be predictive of AI system behaviour in other contexts.

investment. These properties then enable us to outline a concrete path toward a more rigorous evaluation ecosystem for generative AI systems.

2 Lessons from Other Fields

The bridges we stand on, the medicine we take, and the food we eat are all the result of rigorous assessment. In fact, it is because of the rigor of the corresponding evaluation ecosystems that we can trust that the products and critical infrastructure surrounding us are performant and safe. Generative AI products are no exception to this reality and therefore not unique in their need for robust evaluations. In response to their own crises, more established evaluation regimes emerged in other fields to assure users and regulators of safety and reliability - offering concrete lessons for the AI field (Rismani et al. 2023; Raji and Dobbe 2023; Raji et al. 2022a). We note three key evaluation lessons from these other fields: the targeting of real-world performance, the iterative refinement of measurement approaches, and the establishment of functioning processes and institutions.

2.1 Real-world applicability of metrics

First, it is noteworthy that, historically, evaluation made a difference for safety because it tracked real-world risks. Measuring real-world performance does not mean waiting until risks manifest —- on the contrary, earlier pre-deployment risk detection and evaluation allows for more comprehensive and cheaper mitigations (Collingridge 1982). For example, in clinical trials, strict requirements exist for staged, pre-clinical testing in order to minimize risks to vulnerable patient populations. Similarly, airplanes are first designed and tested through simulations to improve understanding of their performance while minimizing risks to life and material damage.

Pre-deployment testing may help identify real-world risks earlier – however, it must be accompanied by *post-deployment monitoring* to detect emergent harms as they happen. For instance, unexpected side effects and off-label use of pharmaceuticals in the medical domain, especially on under-tested populations, are nearly impossible to anticipate pre-deployment. Many of these issues only emerge from highly complex interactions at the point of use. In such cases, health providers, patients and manufacturers are required to report adverse events to regulatory agencies via incident databases [5]. The collection of these incidents and the resulting analyses can then be used to inform any restriction or cautionary uses of the drug or vaccine, especially for at-risk populations. As an example, the discovery of Myocarditis symptoms from the COVID-19 vaccine was facilitated by the Vaccine Adverse Events System (VAERS) incident database. This finding led to a warning and an adjusted dosage recommendation for the most impacted population of male vaccine recipients, aged 12 to 17 (Oster et al. 2022). In some cases, monitoring data can even be used to feed back into future pre-deployment evaluation practices – for example, the results of race-based failures observed in an FDA incident database for medical devices (FDA MAUDE Database) informed new health department guidelines on adequate equitable representation in pre-clinical trials for such devices (U.S. Food & Drug Administration 2017; Fox-Rawlings et al. 2018).

2.2 Iteratively Refining Metrics

The metrics and measurement approaches of evaluation must be iteratively refined and calibrated over time. This iterative process includes choosing and refining relevant measurement *targets*, i.e. the concepts to be measured. Initially, the automotive industry focused on human-caused errors, responding with drivers' education, drivers' licenses, and laws against drunk driving. However, as accidents continued to soar, seatbelt regulations and other design choices became a focal point, feeding into notions of a car's "crashworthiness" tied to manufacturer responsibility (Díaz and Costas 2020). This measurement target of crashworthiness has continued to evolve over time. For example in Europe, concerns about the safety of pedestrians and cyclists were incorporated in an expanded notion of crashworthiness (United Nations 2011), broadening what it means for a car to be considered "safe".

As a measurement target is refined over time, so are the measurement instruments that are designed to capture it. With the measurement of temperature, divergent thermometer readings revealed the importance of engineering instruments with a reliable liquid indicator (Chang 2001). Further

---

[5] For example: https://open.fda.gov/data/faers/, https://vaers.hhs.gov/, https://yellowcard.mhra.gov.uk/

attempts to calibrate thermometers gave rise to deeper insights about temperature itself – as an indication of matter phase changes (i.e. Celsius), human body responses (i.e. Fahrenheit), and quantum mechanical properties (i.e. Kelvin). However, no single measurement instrument is perfect – by triangulating results from multiple methods, more robust insights can be gained (Campbell and Fiske 1959; Jespersen and Wallace 2017). Ultimately, identifying measurement targets, designing metrics, and developing measurement instruments, are all interdependent tasks that require a careful iterative process.

2.3 Establishing Institutions & Norms

A successful evaluation ecosystem requires investing in institutions. The advocacy of Harvey Wiley, Samuel Hopkins Adams, and others led to the 1938 passing of the United States Federal Food, Drug, and Cosmetic Act. This act led to the creation of the Food and Drug Administration (FDA), an agency that is now widely known for its rigorous pharmaceutical and nutrition testing regimes. At the FDA, Wiley and his team developed numerous innovative methods for identifying the presence and effects of particular poisonous ingredients, notably leading several multi-year experiments to assess the pernicious effects of various chemicals on a group of volunteers known as the "Poison Squad" (Blum 2018). Without the centralization of testing efforts through a single agency, this team could not have had the resources or coordination capacities to execute such long-term and large-scale experiments.

In many fields, readily available evaluation tools, shared evaluation infrastructure, and standards afforded by such institutions have contributed to the establishment of more thorough evaluation regimes (Vedung 2017; Timmermans and Berg 2003). After the number of cars on the road increased by an order of magnitude throughout the early 20th century, the corresponding increase in fatal crashes pushed Ralph Nader and other advocates to establish the National Traffic and Motor Vehicle Safety Act in 1966, responsible for the National Highway Safety Bureau (now the transportation testing agency known as the National Highway Traffic Safety Administration, NHTSA). By 1985, Ralph Nader claimed "programs, which emphasize engineering safety, have saved more than 150,000 lives and prevented or reduced in severity a far larger number of injuries" (Nader 1985). In 2015, a NHTSA report revealed that this trend has continued, with an estimated 613,501 lives saved between 1960 to 2012 (Britannica). Nader attributed much of this success to the meaningful enforcement of government-mandated standards, including active monitoring – i.e., regularly measuring everything from fuel efficiency to auto handling and braking capabilities—by the NHTSA, which led to the recall of millions of defective vehicles and tires by the early 1980s.

3 Towards an Evaluation Science for Generative AI

3.1 Unique Challenges of Generative AI

While drawing on lessons from other fields, it is important to understand what makes the challenge of evaluating generative AI systems unique. Other systems – from personal computers to pharmaceuticals – can be used for purposes that were not originally intended. However, generative AI systems are often explicitly designed to be open-ended – that is, underspecified and deliberately versatile in the range of use cases they support (Hughes et al. 2024). This open-endedness makes it hard to define precise measurement targets in AI evaluation, resulting in vague targets such as the long-standing trend of measuring an AI system's "general intelligence", rather than performance on specific tasks (Raji et al. 2021). Furthermore, generative AI systems tend to be less *deterministic* – i.e., the same input can lead to different outputs due to their stochastic nature, and due to unknown factors in training data (Raji 2021). This non-determinism makes it harder to predict system behaviors compared to prior software systems, as it is difficult to directly trace system design choices – about training data, model design or the user-interface – to downstream system outputs and impacts.

Further adding to the complexity of anticipating and evaluating AI system outputs and use cases is the possibility of longitudinal *social* interactions with generative AI systems. This gives rise to a new class of interaction risks that may evolve in unexpected ways over time (e.g., harmful human–AI "relationships" (Manzini et al. 2024). Taken together, these unique challenges inherent to generative AI systems indicate the need for a *behavioral* approach to evaluating such systems, focusing on AI system performance in the context of different real-world settings (Rahwan et al. 2019; Wagner et al.

2021; Matias 2023). Indeed, adopting a behavioral approach that treats AI systems as blackboxes can be helpful in enabling some translation between higher-level systemic impact evaluations and lower-level computational methods (McCoy et al. 2024; Shiffrin and Mitchell 2023).

3.2   Real-world applicability of metrics

There is a disconnect between the current AI evaluation culture, with its focus on benchmarking models, and real-world, grounded approaches to the assessment of performance and safety (Lazar and Nelson 2023). Addressing this divide will require taking deliberate steps to shift the culture surrounding generative AI evaluations from "basic research" toward "*use-inspired* basic research" (Stokes 1997), where the focus is on advancing our scientific understanding of AI system properties and patterns that are relevant for their performance and safety in real-world deployment contexts.

Evaluations of generative AI systems cannot be "one size fits all." As with other fields, even pre-deployment evaluations need to take real-world deployment contexts into account. This echoes several recent calls for holistic, AI system-focused evaluations that take into account relevant context beyond the scope of the current model-focused evaluation culture (Lum et al. 2024; Goldfarb-Tarrant et al. 2021b; Bommasani and Liang 2024; Saxon et al. 2024; Weidinger et al. 2023). To achieve this, AI evaluation science must employ a range of approaches that can respond to different evaluation goals, and move beyond coarse grained claims of "general intelligence" towards more task-specific and real-world relevant measures of progress and performance (Bowman and Dahl 2021; Raji et al. 2021). A variety of more holistic evaluation methods and instruments, appropriate for differing deployment contexts and evaluation goals, need to be developed (NAIAC 2024; Bommasani and Liang 2024; Solaiman et al. 2024; Weidinger et al. 2023; Dobbe 2022). By December 2023, less than 6% of generative AI evaluations accounted for human–AI interactions, and less than 10% considered broader contextual factors (Rauh et al. 2024).

To account for factors beyond technical specifications that influence real-world performance and safety, generative AI evaluations will need to adopt a broader sociotechnical lens (Selbst et al. 2019; Chen and Metcalf; Wallach et al. 2024). Although there is an emerging interest in other approaches, such as more interactive, dynamic, context-rich, and multi-turn benchmarks (Chiang et al. 2024; Saxon et al. 2024; Zhou et al. 2024; Magooda et al. 2023), large gaps remain. For one, anticipating and understanding real-world risks from sustained, personalized human–AI interactions will require more longitudinal studies than have been published to date (e.g., Lai et al. (2023)) and the establishment of post-deployment monitoring regimes for AI systems (e.g., Feng et al. (2025)). Furthermore, insights from real-world deployment need to feed back into early-stage evaluation design – certain existing efforts, such as Anthropic's Clio (Anthropic 2024a) or AllenAI's WildBench (Lin et al. 2024), indicate some promise toward an approach of developing pre-deployment benchmarks with data from "naturalistic" interactions from post-deployment contexts.

3.3   Iteratively Refining Metrics

Developing an evaluation science for generative AI systems requires first identifying which concepts should be measured -–– that is, to determine the proper measurement targets. Common targets of interest in the AI context are often abstract and even contested (Wallach et al. 2024). Operationally defining metrics that capture these targets involves identifying relevant, tractable subcomponents. Take the widely cited risk of "misinformation": relevant factors include whether factually correct information is being provided, the subtlety of whether different persons are likely to believe that information, and how such information may be uncritically disseminated. Each of these aspects is best measured at different levels of analysis – factual accuracy can be determined based on model output, believability requires human-computer interaction studies, and assessing dissemination pathways requires studying the broader systems into which AI is deployed (Weidinger et al. 2023). Triangulating measurements across these levels of analysis can provide a more holistic picture of "misinformation" propagation.

Better integration of evaluation metrics across AI development and deployment can be used to further refine, calibrate, and validate these metrics, enabling an iterative scaffolding of this evaluation science (Wimsatt 1994). Comparing the results of pre-deployment evaluations, such as

static benchmarks, to post-deployment evaluations and monitoring enables an evaluation feedback loop, whereby early-stage evaluations can become better calibrated to take real-world deployment contexts into account. For example, comparing results from static benchmark testing and post-deployment monitoring, one might identify that some AI generated computer code is functional, but frequently misunderstood and falsely applied by users. This insight can then be used to improve benchmarks and other early-stage model testing protocols – e.g. by adopting tests to assess code *legibility*, in addition to testing the functionality of produced computer code (Nguyen et al. 2024).

3.4 Establishing Institutions & Norms

A successful evaluation ecosystem requires investment. Current infrastructure falls short of the systematic approach and effectiveness of evaluation regimes in other fields, where evaluation processes are more costly, complex, and distributed between different actors and skill sets (Raji et al. 2022b; Anthropic 2023; Caliskan and Lum 2024). Prioritizing such investments and developing readily available tools for auditing and evaluation (Ojewale et al. 2024) – including resources to enable the expanded methodological toolkit mentioned above and mechanisms for institutional transparency (White House 2025; Caliskan and Lum 2024) – will be critical in order for AI evaluation practice to become systematized, effective and widespread.

It is already clear that aiming for uncompromised, transparent and open evaluation platforms will come at a significant financial cost. Open source efforts such as Hugging Face's LLM Leaderboard, Eleutheur AI's LLM evaluation harness, Stanford's HELM, and ML Commons provide shared technical infrastructure on which to compare and rank benchmarking results, and there are nascent, but comparable, publicly funded government efforts such as the UK AI Safety Institute's platform *Inspect* and the US National Institute of Standards and Technology pilot of *ARIA* [6]. However, running HELM once on the 30 models assessed in 2022 cost USD $38,000 for the commercial model APIs, and required 20,000 A100 hours of compute to test the open models – even with Anthropic and Microsoft allowing to run their models for free (Liang et al. 2023). This differs glaringly from the cost of running an evaluation on traditional benchmarks such as SQuAD (Rajpurkar et al. 2016) or other GLUE tests (Wang et al. 2019), both of which could be easily downloaded to a personal laptop and run within a few hours at most. Even as specific platforms evolve and expand, this indicates that the next era of evaluation infrastructure for generative AI systems will require financial resources beyond what has been invested so far. Given the history of overlooking the importance of evaluation practices in the machine learning field (Paullada et al. 2020), prioritizing and investing in evaluations will be critical to ensuring safe and trustworthy AI systems.

Shared AI evaluation infrastructure can involve much more than just a community leaderboard. Common AI evaluation tools for everything from harm discovery, standard identification and more can facilitate the evaluation process and provide guidance for evaluation best practice across stakeholders in industry and beyond (Ojewale et al. 2024; Wang et al. 2024). For instance, many documentation efforts provide direct and indirect guidance to engineering teams on how to approach AI evaluation – in order to record the requested information in the template, practitioners must, at minimum, satisfy requirements of a particular evaluation process. For instance, the inclusion of disaggregated evaluations in the Model Card template (i.e. evaluating model performance across different demographic subgroups), increased the practice throughout the machine learning field. AI documentation templates such as Model Cards (Mitchell et al. 2019), SMACTR (Raji et al. 2020), Datasheets for Datasets, (Gebru et al. 2021), and Fact sheets (IBM year), as well as multi-year, multi-stakeholder documentation initiatives like ABOUTML (Raji and Yang 2020) all continue to meaningfully guide current model development and evaluation practice – indeed, several of these documentation templates are being integrated into open-source AI model platforms (Liang et al. 2024), and policy requirements (Kawakami et al. 2024). New documentation frameworks specific to generative AI evaluation have begun to emerge from corporate alliances between generative AI model developers to advance evaluation norms and standards in this context (e.g. Partnership on AI; Frontier Model Forum; MLCommons.

---

[6]See https://ai-challenges.nist.gov/aria and https://inspect.ai-safety-institute.org.uk/.

4 Moving Forward

It is tempting to assume that because generative AI systems are widely used and deployed, they must have been subject to the elaborate safety and performance evaluations that we have come to expect in other fields. Sadly, this is not the case. Because generative AI systems have only recently transitioned from the research landscape to the real world, the current evaluation ecosystem is not yet mature. In many cases, the real-world uses of these systems are still evolving and new application domains are being developed. For many considerations on real-world performance and safety, there are simply no valid, reliable evaluations available yet. Closing this gap requires a deliberate effort to invest in and create an evaluation science for generative AI.

However, evaluations are not neutral. Choosing what and how to evaluate privileges some issues at the cost of others – it is not possible to assess all possible use cases and applications, requiring further prioritisation and value judgement. One principled and responsible approach may be to focus on the highest-risk deployment contexts, such as applications in medicine, law, education and finance – or focus on deployments impacting the most vulnerable populations. A hope may be that by focusing on evaluating generative AI systems safe in these contexts and for these groups, we may lift many boats and build an evaluation ecosystem that makes for more reliable, trustworthy and safe generative AI systems for all.

The trust we have in every product we regularly make use of – from the toaster used to heat our breakfast, to the vehicle mediating our morning commute – has been hard-earned. Valuable insights from safety engineering and measurement science in other fields – such as anticipating real-world failures pre-deployment and monitoring incidents post-deployment, iteratively refining evaluation approaches, and investing in institutions for accessible and robust evaluation ecosystems – can be adopted to mature practices in the AI field. The unique challenges of generative AI technologies do not absolve the field from this responsibility, but further reinforce a clear need for creating an evaluation science it can call its own.

5 Acknowledgements

References

Anthropic. Challenges in evaluating AI systems, October 2023. URL https://www.anthropic.com/research/evaluating-ai-systems.

Anthropic. Clio: Privacy-preserving insights into real-world AI use, December 2024a. URL https://www.anthropic.com/research/clio.

Anthropic. Introducing the next generation of Claude, March 2024b. URL https://www.anthropic.com/news/claude-3-family.

Pete Archer and Oli Elliott. Representation of BBC News content in AI Assistants. Technical report, BBC, February 2025. URL https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf.

Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. AI auditing: The Broken Bus on the Road to AI Accountability, January 2024. URL http://arxiv.org/abs/2401.14462. arXiv:2401.14462.

Deborah Blum. *The Poison Squad: One Chemist's Single-Minded Crusade for Food Safety at the Turn of the Twentieth Century*. Penguin, New York, 2018.

Rishi Bommasani and Percy Liang. Trustworthy Social Bias Measurement. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):210–224, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31630. URL https://ojs.aaai.org/index.php/AIES/article/view/31630.

Samuel R. Bowman and George E. Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding?, October 2021. URL http://arxiv.org/abs/2104.02145. arXiv:2104.02145.

Eric Boyd. Microsoft and Epic expand AI collaboration to accelerate generative AI's impact in healthcare, addressing the industry's most pressing needs, August 2023. URL https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-acce

Britannica. Samuel hopkins adams | biography, works, muckraker, & facts | britannica. URL https://www.britannica.com/biography/Samuel-Hopkins-Adams.

Aylin Caliskan and Kristian Lum. Effective AI regulation requires understanding general-purpose AI, 2024. URL https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purp

Donald T. Campbell and Donald W. Fiske. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi: https://doi.org/10.1037/h0046016.

Hasok Chang. Spirit, air, and quicksilver: The search for the "real" scale of temperature. *Historical Studies in the Physical and Biological Sciences*, 31(2): 249–284, March 2001. ISSN 0890-9997. doi: 10.1525/hsps.2001.31.2.249. URL https://online.ucpress.edu/hsns/article/31/2/249/47889/Spirit-air-and-quicksilver-The-search-f

Brian J. Chen and Jacob Metcalf. A Sociotechnical Approach to AI Policy. URL https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024. URL http://arxiv.org/abs/2403.04132. arXiv:2403.04132.

David Collingridge. *The Social Control of Technology*. Pinter Publishers, London, 1982.

Harm de Vries, Dzmitry Bahdanau, and Christopher Manning. Towards Ecologically Valid Research on Language User Interfaces, July 2020. URL http://arxiv.org/abs/2007.14435. arXiv:2007.14435.

Roel I. J. Dobbe. System Safety and Artificial Intelligence, February 2022. URL http://arxiv.org/abs/2202.09292. arXiv:2202.09292.

Josh Dzieza. The confusing reality of AI friends, December 2024. URL https://www.theverge.com/c/24300623/ai-companions-replika-openai-chatgpt-assistant-romance.

Jacobo Díaz and Miguel Costas. Crashworthiness. In Holm Altenbach and Andreas Öchsner, editors, *Encyclopedia of Continuum Mechanics*, pages 469–486. Springer, Berlin, Heidelberg, 2020. ISBN 9783662557716. doi: 10.1007/978-3-662-55771-6_223. URL https://doi.org/10.1007/978-3-662-55771-6_223.

European Commission. Second Draft of the General-Purpose AI Code of Practice published, written by independent experts, December 2024. URL https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice

Michelle Faverio and Alec Tyson. What the data says about Americans' views of artificial intelligence, November 2023. URL https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of

FDA MAUDE Database. Manufacturer and User Facility Device Experience (MAUDE) Database. URL https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm.

Jean Feng, Fan Xia, Karandeep Singh, and Romain Pirracchio. Not All Clinical AI Monitoring Systems Are Created Equal: Review and Recommendations. *NEJM AI*, 2 (2), January 2025. ISSN 2836-9386. doi: 10.1056/AIra2400657. URL https://ai.nejm.org/doi/10.1056/AIra2400657.

Stephanie R. Fox-Rawlings, Laura B. Gottschalk, Laurén A. Doamekpor, and Diana M. Zuckerman. Diversity in Medical Device Clinical Trials: Do We Know What Works for Which Patients? *The Milbank Quarterly*, 96(3):499–529, September 2018. ISSN 0887-378X. doi: 10.1111/1468-000 9.12344. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131322/`.

Sorelle Friedler, Ranjit Singh, Borhane Blili-Hamelin, Jacob Metcalfe, and Brian J. Chen. AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability, 2023. URL `https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommen`

Frontier Model Forum. Frontier Model Forum: advancing frontier AI safety and security. URL `https://www.frontiermodelforum.org/`.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL `http://arxiv.org/abs/1803.09010`. arXiv:1803.09010.

Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, J. D. Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, and Samer Hassan et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, December 2024. URL `http://arxiv.org/abs/2403.05530`. arXiv:2403.05530.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias, June 2021a. URL `http://arxiv.org/abs/2012.15859`. arXiv:2012.15859.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. Intrinsic Bias Metrics Do Not Correlate with Application Bias, June 2021b. URL `http://arxiv.org/abs/2012.15859`. arXiv:2012.15859.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, and Ehab AlBadawy et al. The Llama 3 Herd of Models, November 2024. URL `http://arxiv.org/abs/2407.21783`. arXiv:2407.21783.

Will Douglas Heaven. Why Meta's latest large language model survived only three days online, November 2022. URL `https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived`

Krystal Hu. ChatGPT sets record for fastest-growing user base - analyst note, February 2023. URL `https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-`

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-Endedness is Essential for Artificial Superhuman Intelligence, June 2024. URL `http://arxiv.org/abs/2406.04268`. arXiv:2406.04268.

IBM. Using AI Factsheets for AI Governance | IBM Cloud Pak for Data as a Service, year. URL `https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/factsheets-model-inventory.ht`

Lone Jespersen and Carol A. Wallace. Triangulation and the importance of establishing valid methods for food safety culture evaluation. *Food Research International*, 100:244–253, October 2017. ISSN 0963-9969. doi: 10.1016/j.foodres.2017.07.009. URL https://www.sciencedirect.com/science/article/pii/S0963996917303319.

Nari Johnson, Elise Silva, Harrison Leon, Motahhare Eslami, Beth Schwanke, Ravit Dotan, and Hoda Heidari. Legacy Procurement Practices Shape How U.S. Cities Govern AI: Understanding Government Employees' Practices, Challenges, and Needs, February 2025. URL http://arxiv.org/abs/2411.04994. arXiv:2411.04994.

Anna Kawakami, Daricia Wilkinson, and Alexandra Chouldechova. Do Responsible AI Artifacts Advance Stakeholder Goals? Four Key Barriers Perceived by Legal and Civil Stakeholders, August 2024. URL http://arxiv.org/abs/2408.12047. arXiv:2408.12047.

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, April 2021. URL http://arxiv.org/abs/2104.14337. arXiv:2104.14337.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1369–1385, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594087. URL https://dl.acm.org/doi/10.1145/3593013.3594087.

Seth Lazar and Alondra Nelson. AI safety on whose terms? *Science*, 381(6654):138–138, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi8982. URL https://www.science.org/doi/10.1126/science.adi8982.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110.

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. What's documented in AI? Systematic Analysis of 32K AI Model Cards, February 2024. URL http://arxiv.org/abs/2402.05160. arXiv:2402.05160.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. August 2021. URL https://openreview.net/forum?id=mPducS1MsEK.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild, October 2024. URL http://arxiv.org/abs/2406.04770. arXiv:2406.04770.

Kristian Lum, Jacy Reese Anthis, Chirag Nagpal, and Alexander D'Amour. Bias in Language Models: Beyond Trick Tests and Toward RUTEd Evaluation, February 2024. URL http://arxiv.org/abs/2402.12649. arXiv:2402.12649 version: 1.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. AI on trial: legal models hallucinate in 1 out of 6 (or more) benchmarking queries, May 2024. URL https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-

Ahmed Magooda, Alec Helyar, Kyle Jackson, David Sullivan, Chad Atalla, Emily Sheng, Dan Vann, Richard Edgar, Hamid Palangi, Roman Lutz, Hongliang Kong, Vincent Yun, Eslam Kamal, Federico Zarfati, Hanna Wallach, Sarah Bird, and Mei Chen. A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications, October 2023. URL `http://arxiv.org/abs/2310.17750`. arXiv:2310.17750.

Arianna Manzini, Geoff Keeling, Lize Alberts, Shannon Vallor, Meredith Ringel Morris, and Iason Gabriel. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):943–957, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31694. URL `https://ojs.aaai.org/index.php/AIES/article/view/31694`.

J. Nathan Matias. Humans and algorithms work together — so study them together. *Nature*, 617(7960):248–251, May 2023. doi: 10.1038/d41586-023-01521-z. URL `https://www.nature.com/articles/d41586-023-01521-z`.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, October 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2322420121. URL `https://pnas.org/doi/10.1073/pnas.2322420121`.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models, October 2024. URL `http://arxiv.org/abs/2410.05229`. arXiv:2410.05229.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting, January 2019. URL `http://arxiv.org/abs/1810.03993`. arXiv:1810.03993.

MLCommons. MLCommons - Better AI for Everyone. URL `https://mlcommons.org/`.

Scott J. Mulligan. The way we measure progress in AI is terrible. URL `https://www.technologyreview.com/2024/11/26/1107346/the-way-we-measure-progress-in-ai-is-terri`

Samantha Murphy Kelly. Microsoft's Bing AI demo called out for several errors | CNN Business, February 2023. URL `https://www.cnn.com/2023/02/14/tech/microsoft-bing-ai-errors/index.html`.

Ralph Nader. Opinion | How law has improved auto technology. *The New York Times*, December 1985. ISSN 0362-4331. URL `https://www.nytimes.com/1985/12/29/opinion/how-law-has-improved-auto-technology.html`.

The National Artificial Intelligence Advisory Committee NAIAC. Findings & recommendations: AI safety. Technical report, May 2024. URL `https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf`.

National Institute of Standards and Technology. AI Risk Management Framework, July 2021. URL `https://www.nist.gov/itl/ai-risk-management-framework`.

Sydney Nguyen, Hannah McLean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. How Beginning Programmers and Code LLMs (Mis)read Each Other, 2024. URL `https://arxiv.org/html/2401.15232v1`.

Katie Nicholson. Bing chatbot says it feels violated and exposed after attack. *CBC News*, February 2023. URL `https://www.cbc.ca/news/science/bing-chatbot-ai-hack-1.6752490`.

Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, March 2024. URL `http://arxiv.org/abs/2402.17861`. arXiv:2402.17861.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *npj Digital Medicine*, 6 (1):1–4, October 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00939-z. URL https://www.nature.com/articles/s41746-023-00939-z.

OpenAI. GPT-4o System Card, 2024. URL https://openai.com/index/gpt-4o-system-card/. publisher:.

Matthew E. Oster, David K. Shay, John R. Su, Julianne Gee, C. Buddy Creech, Karen R. Broder, Kathryn Edwards, Jonathan H. Soslow, Jeffrey M. Dendy, Elizabeth Schlaudecker, Sean M. Lang, Elizabeth D. Barnett, Frederick L. Ruberg, Michael J. Smith, M. Jay Campbell, Renato D. Lopes, Laurence S. Sperling, Jane A. Baumblatt, Deborah L. Thompson, Paige L. Marquez, Penelope Strid, Jared Woo, River Pugsley, Sarah Reagan-Steiner, Frank DeStefano, and Tom T. Shimabukuro. Myocarditis cases reported after mRNA-based COVID-19 vaccination in the US from December 2020 to August 2021. *JAMA*, 327(4):331–40, January 2022. URL https://jamanetwork.com/journals/jama/fullarticle/2788346.

Partnership on AI. PAI's Guidance for Safe Foundation Model Deployment. URL https://partnershiponai.org/modeldeployment/.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research, December 2020. URL http://arxiv.org/abs/2012.05345. arXiv:2012.05345.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568 (7753):477–486, April 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1138-y. URL https://www.nature.com/articles/s41586-019-1138-y.

Deborah Raji. The bodies underneath the rubble. In Frederike Kaltheuner, editor, *Fake AI*. Meatspace Press, 2021.

Inioluwa Deborah Raji and Roel Dobbe. Concrete Problems in AI Safety, Revisited, December 2023. URL http://arxiv.org/abs/2401.10899. arXiv:2401.10899.

Inioluwa Deborah Raji and Jingying Yang. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles, January 2020. URL http://arxiv.org/abs/1912.06166. arXiv:1912.06166.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, January 2020. URL http://arxiv.org/abs/2001.00973. arXiv:2001.00973.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark, November 2021. URL http://arxiv.org/abs/2111.15366. arXiv:2111.15366.

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew D. Selbst. The Fallacy of AI Functionality, July 2022a. URL http://arxiv.org/abs/2206.09511. arXiv:2206.09511.

Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, June 2022b. URL http://arxiv.org/abs/2206.04737. arXiv:2206.04737.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, October 2016. URL http://arxiv.org/abs/1606.05250. arXiv:1606.05250.

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1200–1217, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31717. URL https://ojs.aaai.org/index.php/AIES/article/view/31717.

Liz Reid. Generative AI in Search: Let Google do the searching for you, May 2024. URL https://blog.google/products/search/generative-ai-google-search-may-2024/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP models with CheckList, May 2020. URL http://arxiv.org/abs/2005.04118. arXiv:2005.04118.

Shalaleh Rismani, Renee Shelby, Andrew Smart, Edgar Jatho, Joshua Kroll, AJung Moon, and Negar Rostamzadeh. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18, New York, NY, USA, April 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581407. URL https://dl.acm.org/doi/10.1145/3544548.3581407.

Kevin Roose. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*, February 2023. URL https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Kevin Roose. A.I. Has a Measurement Problem. *The New York Times*, April 2024a. URL https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html.

Kevin Roose. Can A.I. Be Blamed for a Teen's Suicide? *The New York Times*, October 2024b. URL https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as Microscopes: A Call for Model Metrology, July 2024. URL http://arxiv.org/abs/2407.16711. arXiv:2407.16711.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 59–68, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

Richard Shiffrin and Melanie Mitchell. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120, March 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2300963120. URL https://pnas.org/doi/10.1073/pnas.2300963120.

Natasha Singer. In Classrooms, Teachers Put A.I. Tutoring Bots to the Test. *The New York Times*, June 2023a. URL https://www.nytimes.com/2023/06/26/technology/newark-schools-khan-tutoring-ai.html.

Natasha Singer. New A.I. Chatbot Tutors Could Upend Student Learning. *The New York Times*, June 2023b. URL https://www.nytimes.com/2023/06/08/business/khan-ai-gpt-tutoring-bot.html.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Canyu Chen, Hal Daumé III, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Ria Kalluri, Alberto Lusoli, Alina Leidinger, Michelle Lin, Xiuzhu Lin, Sasha Luccioni, Jennifer Mickel, Margaret Mitchell, Jessica Newman, Anaelia Ovalle, Marie-Therese Png, Shubham Singh, Andrew Strait, Lukas Struppek, and Arjun Subramonian. Evaluating the Social Impact of Generative AI Systems in Systems and Society, June 2024. URL http://arxiv.org/abs/2306.05949. arXiv:2306.05949.

Donald E. Stokes. *Pasteur's Quadrant*. Brookings Institution Press, Washington, D.C., 1997.

Stefan Timmermans and Marc Berg. *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*. Temple University Press, Philadelphia, PA, 2003.

United Nations. Proposal to develop amendments to global technical regulation No. 9 concerning pedestrian safety., 2011. URL https://unece.org/fileadmin/DAM/trans/main/wp29/wp29wgs/wp29gen/wp29glob/ECE-TRANS-180-Add9-Am

U.S. Food & Drug Administration. Evaluation and Reporting of Age-, Race-, and Ethnicity-Specific Data in Medical Device Clinical Studies, September 2017. URL https://www.fda.gov/regulatory-information/search-fda-guidance-documents/evaluation-and-report

Evert Vedung. *Public policy and program evaluation*. Routledge, New York, 2017.

Pranshu Verma. They fell in love with AI bots. A software update broke their hearts. *The Washington Post*, March 2023. URL https://www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/.

Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866): 197–204, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03666-1. URL https://www.nature.com/articles/s41586-021-03666-1.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Evaluating Generative AI Systems is a Social Science Measurement Challenge, November 2024. URL http://arxiv.org/abs/2411.10939. arXiv:2411.10939.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, February 2019. URL http://arxiv.org/abs/1804.07461. arXiv:1804.07461.

Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating ai fairness. *Patterns*, 5(11), 2024.

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, October 2023. URL http://arxiv.org/abs/2310.11986. arXiv:2310.11986.

White House. Removing Barriers to American Leadership in Artificial Intelligence, January 2025. URL https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-lead

Kyle Wiggers. The AI industry is obsessed with Chatbot Arena, but it might not be the best benchmark, September 2024. URL https://techcrunch.com/2024/09/05/the-ai-industry-is-obsessed-with-chatbot-arena-but-it-might-

William C. Wimsatt. The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume*, 20:207–74, 1994.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A Careful Examination of Large Language Model Performance on Grade School Arithmetic, November 2024. URL http://arxiv.org/abs/2405.00332. arXiv:2405.00332.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, and Maarten Sap. HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions, October 2024. URL http://arxiv.org/abs/2409.16427. arXiv:2409.16427.