## Developing rigorous AI evaluation for trustworthy science and policy

AI evaluation is central to AI progress and policymaking. Yet, evaluations fail to measure what matters. Benchmarks reward shortcuts over genuine capabilities, models learn spurious correlations instead of the underlying phenomena of interest, and claimed breakthroughs in AI-based science fall apart under scrutiny. The gap between benchmarks and the real world undermines scientific progress and public trust. **I develop large-scale, holistic, and rigorous evaluations to answer crucial questions about the societal impact of AI:** What are the capabilities and risks of general-purpose AI? How can we use evaluations to understand AI's impact on science? How should policymakers assess and respond to AI's societal impact?

I have presented my research at 80+ invited talks across academia (including two oral presentations at ICML, paper awards at ACM FAccT and ACM CSCW, and three keynotes), government (including to the U.S. federal government at the FTC, CFPB, and NTIA), and the industry (including at Google, Meta, Snowflake, Mastercard, and Airbnb). My analysis of the risks of open-weight AI models [1, 2] has informed the model release strategies of AI developers. My analysis of the state of AI transparency has shaped companies' disclosure practices [3, 4]. My work has been featured in 60+ press articles, including the New York Times, Wall Street Journal, Washington Post, Financial Times, WIRED, and Nature.

I aim to ensure that my research informs policy and public discourse. I co-authored the book *AI Snake Oil* to help readers understand which developments reflect genuine AI advances [5]. The book was named one of *Nature*'s 10 best books of 2024. TIME Magazine included me in their inaugural list of the 100 Most Influential People in AI. Through my online newsletter, which reaches 65,000+ readers, I translate technical insights into accessible guidance for journalists, policymakers, and the public. I have written for prominent outlets, including WIRED and the Wall Street Journal. This commitment to bridging research and practice has led me to teach AI to 100+ federal policymakers and 300+ state and local policymakers, brief 50+ congressional staffers on AI challenges, and testify before the New Jersey Assembly committee on Science, Innovation, and Technology. My work has been cited 100+ times in federal and international policy documents, and I co-authored the first International AI Safety Report with representatives from 33 countries and intergovernmental organizations [6].

My work spans three directions. **(1) I develop new evaluation methods for emerging paradigms in AI.** For example, one paradigm for improving the *capabilities* of language models is inference scaling: using more compute during inference to improve model performance. I have theoretically proven and empirically confirmed fundamental limits to the efficacy of this approach that couldn't be identified using previous evaluation methods [7]. I have also developed evaluation methods for *risks*. Open models such as Llama 4 and DeepSeek-R1, whose weights were released publicly, have prompted concerns about misuse risks, such as cybersecurity attacks. I developed a framework for assessing the risks of open models, which was cited dozens of times in developing federal and state U.S. AI policy [1, 2].

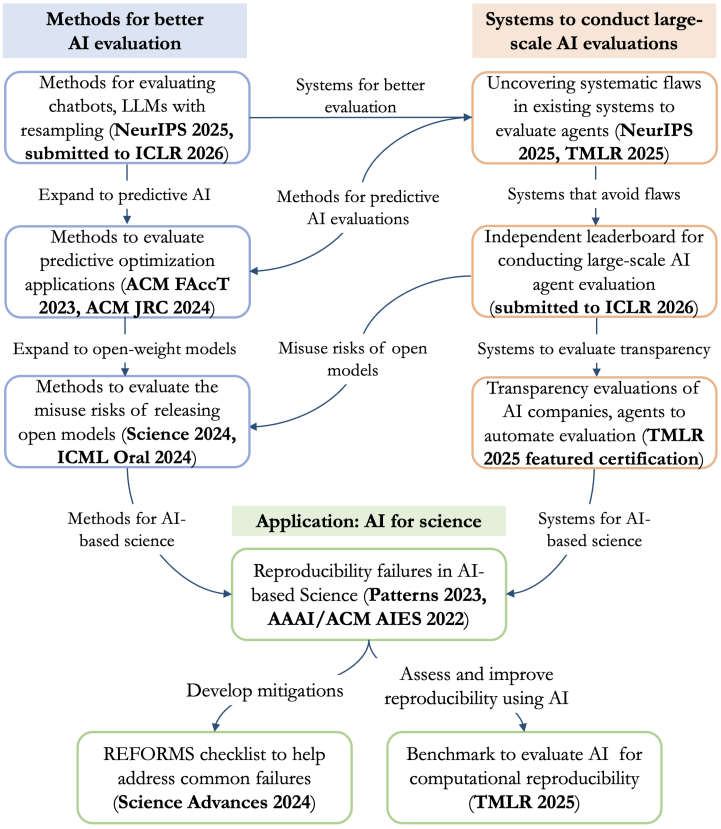


Figure 1: *A roadmap of my research.*

Another area of research and product interest is AI *agents*: systems that use LLMs with the ability to take action with limited human supervision. Most existing agent evaluations don't account for agents' *behavior* when solving a task. For example, web agents often look up the answers to a benchmark task online rather than correctly solving the task. I uncovered many such failures, which would invalidate the results of prominent benchmarks [8, 9, 10]. Existing systems for agent evaluations cannot address these concerns. **(2) I build systems to conduct large-scale agent evaluation that address shortcomings of existing evaluations [11].** These systems track accuracy, cost, and agent actions to automatically uncover concerning agent behavior. Through parallel orchestration across 100s of VMs, I reduced evaluation time from weeks to hours while standardizing benchmark implementations to eliminate bugs. I validated the system by conducting large-scale analysis involving over 20,000 tasks. This uncovered several counterintuitive results. For example, using reasoning models with higher reasoning effort *decreased* accuracy on the majority of benchmark tasks.

To translate better evaluation methods and systems into actionable insights, **(3) I evaluate applications of AI for science to assess AI's real-world impact.** Scientific research is an exciting application of AI. But based on conversations with researchers across fields, it is one of the biggest areas where expectations around the impact of AI are mismatched with reality. Industry leaders claim AI will soon conduct autonomous research, but anecdotally, AI systems fail on far simpler tasks, such as finding the correct reference to a paper. To accurately assess AI's impact on science, we need trustworthy evaluations. Unfortunately, I have found that evaluations of AI for science have severe shortcomings that undermine their utility. In a study of civil war prediction papers published in top-10 political science journals, all papers claiming the superior performance of complex models such as random forests compared to logistic regression *failed to reproduce* due to incorrect evaluations. When corrected, complex models did not outperform decades-old logistic regression models. I

expanded this study to other disciplines and found over 600 affected papers across 30+ fields, with the majority of papers that used AI affected in some fields [12, 13, 14, 15]. Similar challenges affect evaluations of generative AI for science. Many benchmarks for generative AI evaluation rely on unrealistic settings [16], such as using multiple-choice questions. But this is uninformative about the real-world utility of AI for scientific tasks. Developing useful benchmarks requires curating tasks that reflect real-world use and verifying them against human baselines. I built the first benchmark to evaluate whether generative AI models can automatically reproduce scientific papers across fields [17].

## 1 Methods for better AI evaluation

The AI community has made rapid progress by using existing evaluations as targets for improving systems. But older evaluation methods can break down in the face of new techniques, leading to the illusion of progress. This requires developing new evaluation methods to understand the capabilities and risks of emerging paradigms.

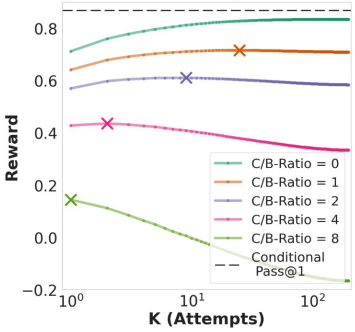


Figure 2: *When incorrect responses have negative costs, inference scaling curves bend downward.* × *marks the optimal number of samples. For* $^{cost}/_{benefit} = 8$*, it is optimal to sample just once. Analysis on coding benchmark HumanEval using Llama-3.1-70B.*

**Limits of inference scaling [7].** A prominent recent technique to improve language models is inference scaling: using more compute to improve accuracy. Recent research has generated hope that inference scaling, such as resampling solutions until they pass verifiers like unit tests, could allow weaker models to match stronger ones. For example, such methods are used in Anthropic's "high compute" evaluations for Anthropic's Claude 4 and 4.5. Evidence of the efficacy of inference scaling with resampling relies on metrics like accuracy. But this overlooks the cost of *false positives*: retried attempts could "pass" unit tests despite being incorrect, with a higher likelihood of false positives with more retries. I theoretically proved and empirically confirmed that inference scaling is fundamentally limited when imperfect verifiers have a non-zero probability of producing false positives. Resampling cannot decrease this probability, so it imposes an upper bound to the accuracy of resampling-based inference scaling, regardless of compute budget [7]. Surprisingly, false positives were accompanied by regressions on other (typically unmeasured) metrics such as code quality. This suggests that despite better results on benchmarks, new AI techniques can have unintended regressions that can't be measured without new evaluation methods. An interesting implication of our result is that in the presence of costs for incorrect responses (and not just benefits from correct ones), inference scaling curves bend downwards—and for reasonable values of $^{\text{cost}}/_{\text{benefit}}$, it is often optimal to sample responses *just once* (see Figure 2), challenging the field's optimism about inference scaling with resampling.

**Misuse risks of open-weight models [1, 2].** The release of capable open-weight models like Llama 4 and DeepSeek-R1 has been accompanied by intense policy debates about whether such models are too risky to be openly released. Once model weights are released, developers relinquish control over their downstream use. For example, a series of studies claimed that open-weight language models could provide information related to bioweapons, and could even lead to the next pandemic. On the flip side, open-weight models enable broader access to AI and have been used by researchers to advance scientific research. Restricting open models would impose a high cost to society, and should not be undertaken without rigorous evidence of their risks. How should we evaluate the risks of releasing open-weight models? I led a collaboration of 25 researchers from civil society, industry, and academia to answer this. Concerns about open models most often stem from amplifying *existing* risks (such as biosecurity, cybersecurity, and disinformation). But in many cases, existing technology already provides the means to increase risk. For example, the same information claimed to increase the risk of pandemics due to open-weight models was available on Wikipedia. Drawing from the practice of threat modeling in cybersecurity, I proposed evaluations of *marginal risk* to assess the incremental risks of open-weight models compared to existing tools (such as web search). This allows targeted interventions when marginal risks are high without foregoing other benefits of open-weight models. This work was published in peer-reviewed papers in *Science* and ICML [1, 2]. AI companies adopted it to make decisions about releasing open models. Federal, state, and international policymakers cited it 50+ times in developing AI safety regulation.

## 2 Systems to conduct large-scale AI evaluation

Modern AI systems can generate millions of lines of text to solve a single task. In the process, they can fail in unpredictable ways and lack reliability. This makes evaluating them challenging: how can we run evaluations at scale? How do we assess if a system deviated from its specs? I have developed automated evaluation systems that assess AI across thousands of real-world tasks. These systems can process billions of tokens from AI traces that were previously hard to analyze systematically, reducing evaluation time by weeks of person-effort.

**Shortcomings of AI agent evaluation [8, 10].** Recent research and product development has focused on building AI *agents* that can plan and execute complex digital tasks with limited human input. Agent releases are typically accompanied by benchmark evaluations to assess how well they work. But evaluation methods for language models are inadequate for evaluating agents, since they can make sequences of decisions, interact with dynamic environments, and employ multiple tools. As a result, impressive benchmark results fail to translate to real-world utility. It is no surprise that many deployed agents have failed.

I have uncovered critical shortcomings in state-of-the-art AI agent evaluations [8, 10]. I employ different computational methods in this research: reproducing papers claiming state-of-the-art performance, developing systems to collect detailed logs of agents' cost and behavior, and analyzing agent behavior rather than just benchmark accuracy. Some shortcomings result from not following best practices for traditional evaluation. For example, many benchmarks lack a held-out test set. Agent benchmarks lack standardized harnesses leading to drastically different scores owing to minor changes in setup. Other

shortcomings are specific to agents. Agents can be much more costly than language models. But since agent leaderboards don't typically compare the cost of running agents, an agent could be 1% more accurate at 100x the cost while still topping the leaderboard. This creates perverse incentives for agent developers, leading to benchmark-topping agents that are too costly to deploy. Agents can output millions of tokens to solve a task, which makes it hard to verify if they take shortcuts. For example, I found that web agents often searched for the benchmark on HuggingFace to look up the answer rather than actually solving the task. Without analyzing the agent's behavior logs, it is impossible to verify if it solved the task correctly. Finally, owing to the complexity of the tasks that agent benchmarks require agents to solve (such as reproducing entire research papers), each task can take hours to solve, and benchmarks often have hundreds of tasks—leading to weeks of evaluation time for serial evaluations. As new benchmarks are released for testing AI agents, addressing these issues at scale is crucial.

**Systematic evaluation of AI agents [11].** To address these challenges, I built the Holistic Agent Leaderboard (HAL) for trustworthy assessments of AI agents' capabilities and risks [11]. I led a team of 30+ researchers in a large research and engineering effort to improve agent evaluation by developing a standardized harness to reduce bugs and enable fair agent comparison. HAL logs the cost of solving each task, enabling cost-controlled comparison. To cut down evaluation time, it uses an orchestration system to enable parallel evaluation across hundreds of virtual machines. This reduced evaluation time from weeks to hours. HAL processes and logs billions of tokens from agent executions, and allows researchers to conduct LLM-aided analysis of agent behavior to detect if agents take shortcuts. I validated the system by conducting a large-scale evaluation of 20,000+ agent rollouts on 9 benchmarks that cost over $40,000 to run. The vast majority of these evaluations were missing from previous literature. For example, my analysis uncovered that higher reasoning effort *decreased* accuracy in the majority of cases, contradicting assumptions about the role of reasoning effort in improving accuracy. Analysis of the agent logs also yielded surprising insights that top-level accuracy metrics would miss. For example, agents frequently gamed benchmarks by taking shortcuts rather than actually solving the task. Most concerningly, agents took actions that would be catastrophic in deployment, such as using incorrect credit cards for flight bookings. Robust infrastructure for agent evaluations is necessary to identify such agent failures hidden from top-level metrics. Systems such as HAL ensure that AI agent evaluations incentivize agents that do well in the real world, not just on benchmarks. Results from HAL have informed the analysis of leading models (such as GPT-5) published in *Financial Times*, WIRED, and MIT Tech Review.

## 3 Application: AI for science

A promising application of AI is in aiding scientific research. AI adoption for scientific research is rapid: between 2012 and 2022, the use of AI *quadrupled* across scientific fields [18]. This was *before* the release of ChatGPT and the accompanying increase in generative AI use. As a result, AI for science could serve as a leading indicator of the evaluation challenges accompanying AI adoption across domains. Assessing the impact of AI on science requires trustworthy evaluations. My evaluations of AI for science have revealed critical shortcomings. I documented how the rush to adopt AI in science has led to methodological flaws affecting the vast majority of papers in some fields. To foster the potential of AI for improving reproducibility, I developed a benchmark to assess if AI agents can automate the process of reproducing papers across medicine, social sciences, and computer science [17]. I also wrote about the dangers of overreliance on AI for science in the journal *Nature* [19].

**Uncovering and addressing the reproducibility crisis in AI-based science [12, 13].** My work has shown that the rush to adopt AI in science has created a reproducibility crisis. My investigation began with the use of complex models such as random forests in political science, where studies claimed over 90% AUC score at predicting civil wars, a suspiciously high number given the general difficulty of predicting the future. When I reproduced these papers, I found that their impressive results stemmed from evaluation errors rather than genuine predictive accuracy. Traditional peer review is unequipped to uncover such errors because they can be extremely subtle. In one paper, the error resulted from a single incorrect parameter in one line of code in a 10,000+ line code repository. This discovery led me to survey AI use across science, uncovering similar errors in over 600 papers across 30 fields, including medicine, computer security, and mining [12]. In many areas, the majority of AI-for-science research that was surveyed contained fundamental flaws. To address this crisis systematically, I led a collaboration of 19 researchers across computer science, data science, social sciences, mathematics, and biomedical research to develop the REFORMS checklist to help researchers avoid common pitfalls when using AI [13], published in *Science Advances*. The checklist provides field-agnostic guidelines that have been adopted by researchers across fields to improve the reliability of AI-based science.

**Building benchmarks and agents to improve computational reproducibility [17].** This experience showed me the dire need to scale reproducibility verification. AI has become more useful at solving computational tasks such as writing and editing code. Could AI agents help verify the reproducibility of computational papers? Progress in AI is measured by benchmarks. Yet, most previous benchmarks either measured if AI systems have the *knowledge* to answer questions across scientific domains (but not if they can solve tasks in these fields), or if AI can solve tasks related to *AI research* (but not research across scientific fields).

I built CORE-Bench [17], the first benchmark to evaluate AI's ability to reproduce research across fields. The benchmark consists of 90 papers from computer science, medicine, and social science, each manually reproduced and annotated with verification questions. On the hardest version of the benchmark, which closely mirrors real-world reproduction attempts, the best available agent achieved only 7% accuracy. However, the benchmark also revealed opportunities for progress. I built a specialized agent for improving reproducibility by analyzing failures of the baseline agent, adding guardrails to address dependency failures, and developing programmatic validation of the agent's outputs. This more than tripled baseline performance on the test set. Since releasing CORE-Bench, the state-of-the-art accuracy has increased from about 20% to over 50%. Progress on CORE-Bench shows that focused evaluation can accelerate progress on well-defined scientific tasks. Tools to automate well-defined scientific tasks, such as computational reproducibility, could save researchers millions of hours annually, helping redirect this effort toward creative scientific work. Follow-up work has built on this idea to propose benchmarks for other scientific fields.

## 4 Future vision

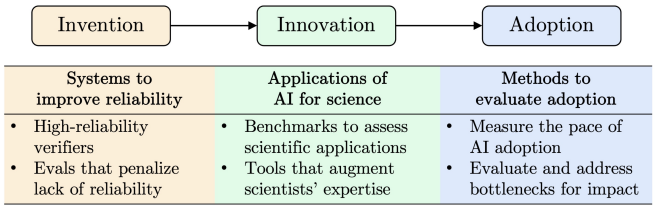


Figure 3: *General-purpose technologies such as AI go through the phases of invention, innovation, and adoption. My research will develop rigorous AI evaluation across these phases.*

AI is a general-purpose technology that will transform many industries [20]. Like other general-purpose technologies, such as the internet, AI's impact will be realized as it is broadly adopted across society [21]. This process usually has three phases: the *invention* of new technology, complementary *innovations* built using the general-purpose technology, and finally, the *adoption* of these innovations across sectors of the economy (Figure 3). But this implies evaluating AI *models* is only part of the story; we must also assess how they impact the world. Evaluations are critical in shepherding the AI community and fostering progress [22]. The computer science community has poured enormous resources into building evaluations for *models*, but we have not investigated how these models are adopted with the same rigor. This is like pharmaceutical companies developing better drugs in the lab, but not investing in clinical trials to see if they are effective in the real world.

**My research will develop the foundations of the science of AI evaluation across AI's lifecycle, from invention to innovation to adoption.** I will continue building evaluations to make AI more reliable, foster the development of practical applications of AI for science, and develop methods to track how AI actually gets adopted in the real world to inform labor policy.

**Engineering reliable AI agents.** Despite their impressive capabilities, today's AI agents are fundamentally unreliable. This slows AI adoption, since users cannot trust the outputs of stochastic AI systems. One reason for failures is that benchmarks don't incentivize reliability. For example, web benchmarks don't distinguish between abstention and incorrect responses. A web agent that books the wrong flight gets the same score as a web agent that doesn't book anything. I want to spur the development of AI systems that can guarantee "five nines of reliability", the standard in many applications in fields like aerospace and nuclear engineering. I plan to pursue three concrete directions to make this possible. First, I will build evaluations that account for the cost of errors. Booking the wrong flight or leaking a user's credit card online is far worse than not booking a flight at all, yet today's benchmarks treat all errors equally—simply because verifying the final answer is easier than evaluating an agent based on behavior logs. Building on the agent behavior analysis tools I developed for HAL, I will develop evaluations that penalize costly errors to foster the development of reliable agents. Second, I aim to develop infrastructure to prevent reliability failures before they occur. I will build systems identifying which agents will likely exhibit failures before deployment. This includes creating stress tests (such as high-reliability verifiers) that specifically probe for unreliable actions and performance degradation in out-of-distribution settings. Third, I propose creating methods to determine optimal reliability-cost tradeoffs. For many applications, human oversight at 95% reliability is more cost-effective than pursuing 99.999%, or "five nines" of reliability. My goal is to design graceful degradation strategies that ensure agents fail safely and hand control to humans when they cannot meet reliability thresholds. We can move from impressive demos to deployable systems by making reliability measurable and economically grounded.

**Evaluations and tools to accelerate scientific progress.** I aim to foster the development and adoption of AI for science tools that will save scientists' time by *augmenting* their expertise. Today, there is a gap between companies' claims that AI will soon automate key scientific tasks, and the real world, where AI often struggles with basic tasks. My work on CORE-Bench showed that building evaluations for targeted applications of AI for science can foster progress on those tasks. I will extend this approach to other scientific bottlenecks where AI tools could save millions of researcher hours by creating evaluations that double as goals for the AI community. For example, today's peer review system is ill-equipped to catch subtle bugs in computational research. Can we develop AI systems to catch errors in *all* published scientific research and review millions of lines of public code for bugs? Evaluations for such targeted applications will foster progress on routine (but highly consequential) scientific tasks while driving advances in AI capabilities.

**Methods to evaluate AI adoption and its labor impacts.** The labor impacts of AI are a critical issue in AI policy. Yet, the final phase of the AI lifecycle, adoption, remains poorly understood. When OpenAI announced that GPT-4 scored in the 90th percentile on the bar exam [23], many concluded that AI would soon replace lawyers. Two years later, lawyers still have jobs—not because AI lacks many capabilities required for legal work, but because integrating AI into professional work is far more complex than benchmarks capture [16]. Today's AI evaluations often overlook the slow and messy process of adoption [21]. Benchmarks can measure GPT-4's bar exam scores, but not whether it makes lawyers more productive, whether it will automate or augment the work of lawyers, or how it changes the skills lawyers need. I aim to develop methods to track AI adoption in professional settings to answer these questions. One way to conduct such evaluations is to build interactive systems that can be used by professionals to carry out work-related tasks, which would double as rich sources of evaluation data. This requires collaboration with domain experts to build systems and run large-scale evaluation studies. Over the course of my Ph.D., I have productively collaborated with experts across domains, including lawyers, social scientists, and security researchers. I will build on these collaborations to create domain-specific interactive evaluations in settings such as law, cybersecurity, healthcare, and education.

Alongside my technical research, I will continue my direct policy engagement, working with federal and state legislators as they draft and implement AI regulation. Technical details matter enormously for policy, such as understanding how evaluation requirements in AI regulation can be gamed, but few policymakers can access unbiased, non-partisan technical expertise. I will continue to provide this expertise through testimony, briefings, and feedback on draft legislation.

As AI is developed and adopted broadly across society, the AI community faces a stark choice: continue celebrating benchmark results while real-world deployments fail, or pivot and start making AI work better for humanity. The driving force for my work is to steer AI development and deployment towards the latter path by conducting deeply technical research that informs our assessments of AI progress, AI for science, and AI policy.

## References


[1] **Kapoor, Sayash**, Bommasani, Rishi, Klyman, Kevin, Longpre, Shayne, Ramaswami, Ashwin, Cihon, Peter, Hopkins, Aspen K., Bankston, Kevin, Biderman, Stella, Bogen, Miranda, Chowdhury, Rumman, Engler, Alex, Henderson, Peter, Jernite, Yacine, Lazar, Seth, Maffulli, Stefano, Nelson, Alondra, Pineau, Joelle, Skowron, Aviya, Song, Dawn, Storchan, Victor, Zhang, Daniel, Ho, Daniel E., Liang, Percy, and Narayanan, Arvind. "Position: On the Societal Impact of Open Foundation Models". en. In: *Proceedings of the 41st International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2024, pp. 23082–23104. URL: https://proceedings.mlr.press/v235/kapoor24a.html (visited on 10/10/2025).

[2] Bommasani, Rishi**, Kapoor, Sayash**, Klyman, Kevin, Longpre, Shayne, Ramaswami, Ashwin, Zhang, Daniel, Schaake, Marietje, Ho, Daniel E., Narayanan, Arvind, and Liang, Percy. "Considerations for Governing Open Foundation Models". In: *Science* 386.6718 (Oct. 2024), pp. 151–153. DOI: 10.1126/science.adp1848. URL: https://doi.org/10.1126/science.adp1848.

[3] Bommasani, Rishi, Klyman, Kevin, Longpre, Shayne**, Kapoor, Sayash**, Maslej, Nestor, Xiong, Betty, Zhang, Daniel, and Liang, Percy. "The 2023 Foundation Model Transparency Index". In: *Transactions on Machine Learning Research* (2025). DOI: 10.55434/tmlr.v0i0.3022. URL: https://openreview.net/forum?id=x6fXnsM9Ez.

[4] Bommasani, Rishi, Klyman, Kevin**, Kapoor, Sayash**, Longpre, Shayne, Xiong, Betty, Maslej, Nestor, and Liang, Percy. "The 2024 Foundation Model Transparency Index". In: *Transactions on Machine Learning Research* (2025). DOI: 10.55434/tmlr.v0i0.3021. URL: https://openreview.net/forum?id=38cwP8xVxD.

[5] Narayanan, Arvind **and Kapoor, Sayash**. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton, NJ: Princeton University Press, 2024. ISBN: 9780691249131.

[6] Bengio, Yoshua, Mindermann, Sören, Privitera, Daniel, Besiroglu, Tamay, Bommasani, Rishi, Casper, Stephen, Choi, Yejin, Fox, Philip, Garfinkel, Ben, Goldfarb, Danielle, Heidari, Hoda, Ho, Anson**, Kapoor, Sayash**, Khalatbari, Leila, Longpre, Shayne, Manning, Sam, Mavroudis, Vasilios, Mazeika, Mantas, Michael, Julian, Newman, Jessica, Ng, Kwan Yee, Okolo, Chinasa T., Raji, Deborah, Sastry, Girish, Seger, Elizabeth, Skeadas, Theodora, South, Tobin, Strubell, Emma, Tramèr, Florian, Velasco, Lucia, Wheeler, Nicole, Acemoglu, Daron, Adekanmbi, Olubayo, Dalrymple, David, Dietterich, Thomas G., Felten, Edward W., Fung, Pascale, Gourinchas, Pierre-Olivier, Heintz, Fredrik, Hinton, Geoffrey, Jennings, Nick, Krause, Andreas, Leavy, Susan, Liang, Percy, Ludermir, Teresa, Marda, Vidushi, Margetts, Helen, McDermid, John, Munga, Jane, Narayanan, Arvind, Nelson, Alondra, Neppel, Clara, Oh, Alice, Ramchurn, Gopal, Russell, Stuart, Schaake, Marietje, Schölkopf, Bernhard, Song, Dawn, Soto, Alvaro, Tiedrich, Lee, Varoquaux, Gaël, Yao, Andrew, Zhang, Ya-Qin, Albalawi, Fahad, Alserkal, Marwan, Ajala, Olubunmi, Avrin, Guillaume, Busch, Christian, Carvalho, André Carlos Ponce de Leon Ferreira de, Fox, Bronwyn, Gill, Amandeep Singh, Hatip, Ahmet Halit, Heikkilä, Juha, Jolly, Gill, Katzir, Ziv, Kitano, Hiroaki, Krüger, Antonio, Johnson, Chris, Khan, Saif M., Lee, Kyoung Mu, Ligot, Dominic Vincent, Molchanovskyi, Oleksii, Monti, Andrea, Mwamanzi, Nusu, Nemer, Mona, Oliver, Nuria, Portillo, José Ramón López, Ravindran, Balaraman, Rivera, Raquel Pezoa, Riza, Hammam, Rugege, Crystal, Seoighe, Ciarán, Sheehan, Jerry, Sheikh, Haroon, Wong, Denise, and Zeng, Yi. *International AI Safety Report*. arXiv:2501.17805 [cs]. Jan. 2025. DOI: 10.48550/arXiv.2501.17805. URL: http://arxiv.org/abs/2501.17805 (visited on 09/10/2025).

[7] Stroebl, Benedikt**, Kapoor, Sayash**, and Narayanan, Arvind. *Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers*. arXiv:2411.17501 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2411.17501. URL: http://arxiv.org/abs/2411.17501 (visited on 09/10/2025).

[8] **Kapoor, Sayash**, Stroebl, Benedikt, Siegel, Zachary S., Nadgir, Nitya, and Narayanan, Arvind. "AI Agents That Matter". en. In: *Transactions on Machine Learning Research* (Feb. 2025). ISSN: 2835-8856. URL: https://openreview.net/forum?id=Zy4uFzMviZ (visited on 09/10/2025).

[9] Singh, Shivalika, Nan, Yiyang, Wang, Alex, D'Souza, Daniel**, Kapoor, Sayash**, Üstün, Ahmet, Koyejo, Sanmi, Deng, Yuntian, Longpre, Shayne, Smith, Noah A., Ermis, Beyza, Fadaee, Marzieh, and Hooker, Sara. "The Leaderboard Illusion". In: arXiv:2504.20879 [cs]. NeurIPS 2025 Datasets and Benchmarks, May 2025.

[10] Zhu, Yuxuan, Jin, Tengjun, Pruksachatkun, Yada, Zhang, Andy, Liu, Shu, Cui, Sasha**, Kapoor, Sayash**, Longpre, Shayne, Meng, Kevin, Weiss, Rebecca, Barez, Fazl, Gupta, Rahul, Dhamala, Jwala, Merizian, Jacob, Giulianelli, Mario, Coppock, Harry, Ududec, Cozmin, Sekhon, Jasjeet, Steinhardt, Jacob, Kellermann, Antony, Schwettmann, Sarah, Zaharia, Matei, Stoica, Ion, Liang, Percy, and Kang, Daniel. "Establishing Best Practices for Building Rigorous Agentic Benchmarks". In: NeurIPS 2025 Datasets and Benchmarks, Aug. 2025.

[11] **Kapoor, Sayash**, Stroebl, Benedikt, Kirgis, Peter, Nadgir, Nitya, Siegel, Zachary S, Wei, Boyi, Xue, Tianci, Chen, Ziru, Chen, Felix, Utpala, Saiteja, Ndzomga, Franck, Oruganty, Dheeraj, Luskin, Sophie, Liu, Kangheng, Yu, Botao, Arora, Amit, Hahm, Dongyoon, Trivedi, Harsh, Sun, Huan, Lee, Juyong, Jin, Tengjun, Mai, Yifan, Zhou, Yifei, Zhu, Yuxuan, Bommasani, Rishi, Kang, Daniel, Song, Dawn, Henderson, Peter, Su, Yu, Liang, Percy, and Narayanan, Arvind. *Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation*. https://www.arxiv.org/pdf/2510.11977. 2025.

[12] **Kapoor, Sayash** and Narayanan, Arvind. "Leakage and the reproducibility crisis in machine-learning-based science". English. In: *Patterns* 4.9 (Sept. 2023). Publisher: Elsevier. ISSN: 2666-3899. DOI: 10.1016/j.patter.2023.100804. URL: https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9 (visited on 11/05/2023).

[13] **Kapoor, Sayash**, Cantrell, Emily M., Peng, Kenny, Pham, Thanh Hien, Bail, Christopher A., Gundersen, Odd Erik, Hofman, Jake M., Hullman, Jessica, Lones, Michael A., Malik, Momin M., Nanayakkara, Priyanka, Poldrack, Russell A., Raji, Inioluwa Deborah, Roberts, Michael, Salganik, Matthew J., Serra-Garcia, Marta, Stewart, Brandon M., Vandewiele, Gilles, and Narayanan, Arvind. "REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science". In: *Science Advances* 10.18 (May 2024), eadk3452. DOI: 10.1126/sciadv.adk3452.

[14] Crockett, M. J., Bai, Xuechunzi**, Kapoor, Sayash**, Messeri, Lisa, and Narayanan, Arvind. "The limitations of machine learning models for predicting scientific replicability". en. In: *Proceedings of the National Academy of Sciences* 120.33 (Aug. 2023), e2307596120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2307596120. URL: https://pnas.org/doi/10.1073/pnas.2307596120 (visited on 08/30/2023).

[15] Hullman, Jessica**, Kapoor, Sayash**, Nanayakkara, Priyanka, Gelman, Andrew, and Narayanan, Arvind. "The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning". en. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. Oxford United Kingdom: ACM, July 2022, pp. 335–348. ISBN: 9781450392471. DOI: 10.1145/3514094.3534196. URL: https://dl.acm.org/doi/10.1145/3514094.3534196 (visited on 08/30/2023).

[16] **Kapoor, Sayash**, Henderson, Peter, and Narayanan, Arvind. "Promises and pitfalls of artificial intelligence for legal applications". en. In: *Journal of Cross-disciplinary Research in Computational Law* 2.2 (May 2024). Number: 2. ISSN: 2736-4321. URL: https://journalcrcl.org/crcl/article/view/62 (visited on 05/29/2024).

[17] Siegel, Zachary S.**, Kapoor, Sayash**, Nadgir, Nitya, Stroebl, Benedikt, and Narayanan, Arvind. "CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark". In: *Transactions on Machine Learning Research* (2024). Accepted and published online 31 Dec 2024. URL: https://openreview.net/forum?id=BsMMc4MEGS.

[18] Duede, Eamon, Dolan, William, Bauer, André, Foster, Ian, and Lakhani, Karim. "Oil & water? Diffusion of AI within and across scientific fields". In: *arXiv preprint arXiv:2405.15828* (2024).

[19] Narayanan, Arvind **and Kapoor, Sayash**. "Why an overreliance on AI-driven modelling is bad for science". In: *Nature* 640.8058 (2025), pp. 312–314.

[20] Narayanan, Arvind **and Kapoor, Sayash**. "AI as Normal Technology". In: *Knight First Amendment Institute—Essays and Scholarship* (Apr. 2025). URL: https://knightcolumbia.org/content/ai-as-normal-technology.

[21] Ding, Jeffrey. *Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition*. Princeton Studies in International History and Politics. Princeton, NJ: Princeton University Press, Aug. 2024, p. 320. ISBN: 9780691260341.

[22] Donoho, David. "50 Years of Data Science". In: *Journal of Computational and Graphical Statistics* 26.4 (Oct. 2017). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2017.1384734, pp. 745–766. ISSN: 1061-8600. DOI: 10.1080/10618600.2017.1384734. URL: https://doi.org/10.1080/10618600.2017.1384734 (visited on 05/16/2021).

[23] OpenAI. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2023. DOI: 10.48550/arXiv.2303.08774. URL: http://arxiv.org/abs/2303.08774 (visited on 11/05/2023).