

RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Governance

Patricia Paskov^{1*}, Kevin Wei¹, Shen Hong², Dan Bateyko³,
Xavier Roberts-Gaal⁴, Carson Ezell¹, Gailius Praninskas⁵, Valerie Chen⁶,
Umang Bhatt⁷, Ella Guest¹

¹RAND, Santa Monica, CA 90407, United States

²Johns Hopkins University, Baltimore, MD 21218, United States

³Cornell University, Ithaca, NY 14853, United States

⁴Harvard University, Cambridge, MA 02138, United States

⁵London School of Economics, London WC2A 2AE, United Kingdom

⁶Carnegie Mellon University, Pittsburgh, PA 15213, United States

⁷University of Cambridge, Cambridge CB2 1TN, United Kingdom

Abstract

Human uplift studies, or studies that measure the effects of AI access on human performance via randomized controlled trials (RCTs) or similar methodologies, increasingly inform frontier AI governance and deployment decisions. While RCT methods are robust in other fields, their interaction with the distinctive properties of frontier AI systems remains underexamined, particularly when results are used to inform high-stakes decisions. We present findings from interviews with 16 expert practitioners with experience conducting human uplift studies in domains including biosecurity, cybersecurity, education, and labor. Across interviews, experts described a recurring tension between the standard causal inference assumptions upon which human uplift studies rely and the object of study itself. Rapidly evolving AI systems, shifting baselines, heterogeneous and changing user proficiency, and porous real-world settings strain assumptions underlying internal, external, and construct validity, complicating the interpretation and appropriate use of uplift evidence. We contribute (1) a synthesis of methodological challenges in human uplift studies, mapped to risks to study validity and classified by their degree of specificity to large language model (LLM) systems, and (2) a mapping from challenges to proposed solutions. By collating expert-identified challenges and solutions, we seek to clarify the interpretive limits and appropriate uses of human uplift evidence, to align evaluation practice with the decisions it informs, and to support more coordinated methodological foundations for AI governance.

Full Version: — <https://arxiv.org/abs/2603.11001>

1 Introduction

Large language model (LLM) systems are reshaping fundamental aspects of human society, from how we work and learn to how we make decisions about national security and public policy (Bengio et al. 2026; Maslej et al.

2025). Society’s ability to anticipate, manage, and govern these transformations depends in part on how well we can evaluate AI’s real-world impacts. Governance frameworks more and more rely on evaluations as a precondition for responsible deployment: the EU AI Act mandates evaluations of frontier models (European Parliament and Council of the European Union 2024); US state-level laws such as New York’s RAISE Act and California’s SB 53 require public reporting of safety evaluations (New York State Assembly 2025; California State Legislature 2025); and frontier developers voluntarily commit to evaluation-based risk thresholds through Responsible Scaling Policies and Frontier Safety Frameworks (e.g., Anthropic 2023; Google DeepMind 2024; OpenAI 2025). Yet the evaluation methods on which this paradigm relies largely compare AI systems with each other and often fall short in measuring how systems affect users and society in practice (Schwartz et al. 2025; Ibrahim et al. 2025; Weidinger et al. 2023; Strauss et al. 2025).

Human uplift studies offer a way to close this gap. By measuring the extent to which access to or use of LLM systems shifts human performance on a task relative to a counterfactual through randomized controlled trial (RCT) or similar methodologies, uplift studies provide rigorous causal evidence about AI’s effects on the people who interact with it (Frontier Model Forum 2025b; UK AI Security Institute 2024a). Human uplift studies are increasingly discussed by international consortia (Bengio et al. 2024; Frontier Model Forum 2025b), governments (European Commission 2025; UK AI Security Institute 2024b,a; U.S. AI Safety Institute 2024), developers (Anthropic 2025; Patwardhan et al. 2024; Shah et al. 2025; Shanghai AI Lab et al. 2025), and AI evaluation contexts (McCaslin et al. 2025; Paskov et al. 2025; METR 2025; Future of Life Institute 2025; Janjeva et al. 2024; Bucknall et al. 2025; Wilson 2024; Karnofsky 2024).¹

*Corresponding author: ppaskov@rand.org
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹The term “human uplift study” has emerged as a consensus label within AI policy and evaluation discourse to describe this class of methods (Bengio et al. 2024; Frontier Model Forum 2025b;

	Methodological Challenges (by Research Phase)								
	Design		Recruitment		Execution		Documentation		
	C1 Measurement [C] ★	C2 Controls [C] ★★	C3 Recruitment [E] ★	C4 AI Literacy [I, E] ★★	C5 Intervention Fidelity [C, I] ★★★	C6 Interference [I] ★★	C7 Expectancy Effects [I, E] ★★	C8 Documentation [E] ★	C9 Interpretation [E] ★
Solutions									
S1 Standardized Task Libraries	✓							✓	✓
S2 Baseline & Control Conventions		✓					✓	✓	✓
S3 Leveling AI Literacy				✓				✓	✓
S4 Versioned Snapshots					✓			✓	✓
S5 Interference Management						✓		✓	✓
S6 Natural Experiments & Phased Rollouts			✓						✓
S7 AI-Accelerated Research Methods			✓	✓	✓	✓			✓
S8 Post-Hoc Analysis			✓	✓	✓	✓	✓		✓
S9 Information Security Frameworks								✓	✓

Table 1: Methodological challenges and solutions in human uplift studies: a mapping between proposed solutions (rows) and methodological challenges (columns) identified in 16 expert interviews. Challenges are grouped by the primary research phase in which they arise, tagged by the primary form of validity they threaten (construct [C], internal [I], external [E]), and rated for LLM-specificity (from familiar RCT threats intensified by LLMs (*), to those with RCT precedent but inadequate standard mitigations (**), to challenges native to LLM-based systems (***)). Mapping is illustrative rather than exhaustive.

This promise, however, outpaces methodological understanding. While RCT methodology is well-established in fields including medicine, economics, and human-computer interaction (Eble, Boone, and Elbourne 2017; Tong et al. 2023; Dahlinger et al. 2018; Jones et al. 2025), its application to frontier AI systems poses new challenges. Moreover, study results are not always made public; even when published, results often omit key methodological details due to safety concerns, such as in Anthropic (2025). As uplift studies increasingly shape consequential decisions about AI deployment, safety, and regulation, gaps between the methodological foundations of this evidence and the institutional conditions under which it is produced and used become questions not only of research design but of AI governance.

This article fills that gap with findings from 16 interviews of experts who have conducted or were contemporaneously conducting human uplift studies involving large language models (LLMs). Our main contributions are: (1) a

UK AI Security Institute 2024a). For clarity, we use “human uplift study” to refer broadly to RCT-based or quasi-experimental evaluations of AI’s effects on human task performance, while acknowledging that individual studies may vary in design details and rigor. We note that the term itself is a misnomer: not all such studies find that access to frontier AI systems increases human performance, e.g., Becker et al. (2025).

synthesis of methodological challenges in LLM uplift studies, mapped to threats to construct, internal, and external validity and classified by their degree of LLM-specificity; and (2) a mapping from challenges to expert-proposed solutions. Our focus and analysis are centered on human uplift studies involving LLM-based systems, though we occasionally use the broader term “AI systems” for generality. By collating expert-identified challenges and solutions, we seek to clarify the interpretive limits and appropriate uses of human uplift evidence, and to promote the development of increasingly rigorous and robust AI evaluations for governance.

2 Background & Related Work

Evaluations provide stakeholders with critical insights into the capabilities, risks, and opportunities of frontier AI systems (Bengio et al. 2026; Burden et al. 2025; Anthropic 2025; Anderson-Samways et al. 2024; UK AI Security Institute 2024a; European Commission 2025). Current approaches, including multiple-choice question-answer (MCQA) benchmarks, red-teaming, and long-form agent evaluations, present distinct trade-offs in validity, reproducibility, and resource requirements (Mouton, Lucas, and Guest 2024; UK AI Security Institute 2024a; Frontier Model Forum 2025b; Reuel et al. 2024; Lin et al. 2024; Burden

et al. 2025; Wei et al. 2025a). While MCQA benchmarks provide structured performance measurement, they often neglect system interaction with users or environments (Weidinger et al. 2023; Schwartz et al. 2025; Ibrahim et al. 2025; Strauss et al. 2025). As such, benchmarks alone poorly predict downstream impacts, especially on economically or strategically important tasks (Schaeffer et al. 2025; Brooks 2025; Narayanan and Kapoor 2025; Ho and Denain 2025). Red-teaming approaches, while involving human-computer interaction, often lack the controlled structure for reliable causal estimates (Feffer et al. 2025; Friedler et al. 2023).

Human uplift studies use RCT or similar methodology to measure the causal impacts of AI systems on human performance, drawing upon decades of experimental rigor from medicine, economics, and social science (Ibrahim et al. 2025; Dobbe, Gilbert, and Mintz 2021; Weidinger et al. 2023; U.S. AI Safety Institute 2024; Farrell et al. 2025; Eble, Boone, and Elbourne 2017; Frontier Model Forum 2025b; Ho and Berg 2025). RCTs originated in medical research as the gold standard for establishing causal relationships between interventions and outcomes, with rigorous standards developed over decades to minimize bias and ensure reliable inference (Eble, Boone, and Elbourne 2017). RCT methodologies have been adopted across fields including human-computer interaction (Tong et al. 2023; Dahlinger et al. 2018; Price et al. 2020; Jeong et al. 2018; Milli, Belli, and Hardt 2022; Lee et al. 2025; Jakesch et al. 2019; G. Mitchell et al. 2021; Bassen et al. 2020) and economics and social sciences (Banerjee 2020; Banerjee, Duflo, and Kremer 2020; Angrist and Pischke 2009). Across these domains, RCTs are grounded in the *Potential Outcomes Framework*, which specifies the assumptions required for causal inference and unbiased treatment-effect estimation (Rubin 1974; Imbens and Rubin 2015).

In recent years, researchers have leveraged RCTs and other methods to evaluate the impact of LLMs on human performance across a growing set of domains, including biological threats (Mouton, Lucas, and Guest 2024; Patwardhan et al. 2024; Krishna et al. 2025; xAI 2025; Zhang et al. 2026), developer productivity (Becker et al. 2025; Chen et al. 2025), legal services (Schwartz et al. 2025; Choi and Schwarcz 2024), customer service (Brynjolfsson, Li, and Raymond 2025), and academic research (Brodeur et al. 2025; Ratkovic et al. 2025). This breadth spans both benefit-oriented questions, such as productivity gains, and worst-case misuse risks. As decision-makers and policymakers increasingly turn to human uplift studies to inform deployment and governance decisions, the methodological soundness of these studies carries more weight.

The application of RCTs to LLMs calls into question the core assumptions of the *Potential Outcomes Framework* (Rubin 1974; Imbens and Rubin 2015), including stable interventions and well-defined counterfactuals. While violations of these assumptions are documented even in mature fields (Eble, Boone, and Elbourne 2017), LLM-based systems place additional and distinctive pressures on these assumptions. Rapid model iteration can undermine intervention fidelity over the course of a study, while the widespread integration of AI tools into everyday workflows (Menlo Ven-

tures 2025; McKinsey & Company 2025) complicates the specification of controls and the prevention of contamination.²

Amid these methodological questions and an evolving empirical landscape, much of the practical experience of conducting human uplift studies remains undocumented. Many studies are conducted under security constraints or commercial confidentiality (Shevlane and Dafoe 2020; Strauss et al. 2025; Delaney, Guest, and Williams 2024; McCaslin et al. 2025), limiting transparency about design trade-offs, failure modes, and interpretive pitfalls. As a result, stakeholders often rely on human uplift evidence without a clear view of where standard assumptions hold, where they break down, and how researchers adapt in practice. In this article, we seek to fill this transparency gap.

3 Methodology

We employ expert interviews as the primary method for this study. Human uplift studies, when published, often only report high-level results and frequently omit details on methodological trade-offs, execution challenges, and interpretive uncertainties. Expert interviews allow us to surface this tacit knowledge directly, including insights from studies that remain unpublished or delayed due to security or commercial constraints. This project was reviewed and deemed exempt by RAND’s Human Subjects Protection Committee.

3.1 Interview protocol

We conducted semi-structured expert interviews between July and August 2025 with experts who had conducted or were, at the time of interview, conducting at least one human uplift study of an LLM-based system. We selected expert participants using a snowball sampling method (Parker, Scott, and Geddes 2019), seeded with a rapid literature review of published human uplift studies involving LLMs that fulfilled the pre-specified inclusion/exclusion criteria. For full sampling procedures, refer to the Appendix.³ In total, we reached out to 53 experts and secured the participation of 16 interviewees (30% participation rate). Interviews were scheduled for 60 minutes and lasted 30–70 minutes. The semi-structured script addressed demographics, human uplift study history, methodological challenges and solutions, and open-ended questions (details in the Appendix).

3.2 Thematic analysis

Following interview completion and transcription, we conducted a qualitative thematic analysis (Braun and Clarke 2006) to identify high-level methodological challenges and solutions. We used a two-stage inductive approach, treating coding as an interpretive and reflexive process (Braun and Clarke 2023) to identify themes from the bottom up. The final codebook contained 30 codes across 7 categories (see

²Intervention fidelity refers to whether the treatment actually delivered matches the treatment specified in the study design. In the AI context, this is threatened when models update, safety filters change, or system configurations shift during a study. For more discussion, see C5.

³Available at <https://arxiv.org/abs/2603.11001>

Appendix), which we organized into broader themes (the Results section). Through thematic analysis, we aimed to identify high-level challenges and potential solutions surfaced in expert interviews. Each transcript was independently coded by two annotators. Following established qualitative research practices (McDonald, Schoenebeck, and Forte 2019; Braun et al. 2024; Li et al. 2022; Wei et al. 2024), we did not calculate inter-rater reliability, given our goal to interpret core narratives rather than make objective or predictive claims.

3.3 Validity mapping

We map challenges to three established forms of validity: construct, internal, and external validity (Table 2). This typology derives from Campbell (1957) and subsequent work Campbell and Cook (1979); Cook, Campbell, and Shadish (2002) and complements recent work in AI evaluation that adopts and extends validity frameworks to assess the reliability and generalizability of model evaluations (Salaudeen et al. 2025; Chouldechova et al. 2024).⁴ Mapping was completed independently by two authors (P.P., K.W.), with a third author adjudicating disagreements to reach consensus. Used as an analytic lens, this framework helps to identify when and how inferential claims are at stake, and seeks to foster more well-informed design and interpretation of human uplift studies.

3.4 LLM-specificity mapping

We further map each challenge along a three-level spectrum reflecting the degree to which it emerges due to the distinctive properties of LLM-based systems (Table 3). Mapping was completed independently by two authors (K.W., C.E.), with a third author adjudicating disagreements to reach consensus.

3.5 Limitations

Our interview process was constrained by sample size and by potential sampling and response bias. Given the nascency of LLM human uplift research and the small population of experts in this area, however, our sample of 16 experts is in line with norms in expert interview research in AI ethics (Wei et al. 2024; Wang, Datta, and Dickerson 2025; Rostamzadeh et al. 2022; Mehandru, Robertson, and Salehi 2022; Cheong et al. 2024; Vrijenhoek et al. 2024; Groves et al. 2024; Deng et al. 2022; Madaio et al. 2024) and human-computer interaction (Agrawal et al. 2021; Zhang et al. 2025; Li et al. 2022; Sekwenz, Wagner, and Parkin 2025; Buruk, Matjeka, and Mueller 2023).⁵ Our interview

⁴The Campbell tradition also considers statistical conclusion validity, which we omit because practitioner-reported challenges concerned causal effects rather than statistical estimation.

⁵The population of experts with direct experience conducting LLM human uplift studies is small; in later rounds of our snowball sampling, interviewees frequently identified experts already on our outreach list, suggesting we were approaching the boundary of the reachable expert population. As one expert noted, “the AI uplift world, it’s still kind of small.” Yet content saturation in similar expert studies has been achieved with a dozen or fewer in-

sample reflects broader biases representative of the field: experts were predominantly U.S.-based, male, and academically affiliated, with only one industry representative responding despite multiple outreach attempts. These limitations may affect the generalizability of findings to non-Western, non-English-speaking, or industry contexts. Limited industry representation likely reflects disclosure constraints rather than limited engagement, underscoring the transparency gap motivating this study. Finally, validity and LLM-specificity mappings reflect our structured analytic judgment rather than objective classification, and are intended as interpretive heuristics rather than sharp ontological claims.

4 Results

4.1 Descriptive statistics

We provide some summary statistics about our expert interviewees, as well as the studies that interviewees had conducted and that we discussed in interviews.

Sample: Experts We interviewed 16 experts (Table 4), 13 of whom identified as men and three as women. Educational backgrounds included PhD ($n = 7$), master’s ($n = 6$), or bachelor’s degrees ($n = 3$). Interviewed experts were spread across a range of seniority levels, with 0–5 years of work experience ($n = 3$), 6–10 years of work experience ($n = 7$), 11–15 years of work experience ($n = 3$), and 16+ years of work experience ($n = 3$). Experts were affiliated with universities ($n = 6$), independent research institutions ($n = 7$), government ($n = 1$), an AI company ($n = 1$), and other organization ($n = 1$). Institutions were based in the U.S. ($n = 14$), Germany ($n = 1$), and unspecified ($n = 1$).

Sample: Human Uplift Studies Experts had completed or were, at the time of interview, conducting between 1 and 6 LLM human uplift studies, with most working on just one ($n = 10$). We discussed 16 distinct studies total, including those published ($n = 9$), under review ($n = 4$), and not-expected-to-be-published due to proprietary or security concerns ($n = 3$). Studies focused primarily on biology/biological risk ($n = 6$), with others spanning software engineering, cybersecurity, medicine, social sciences, and other domains. Seven studies aimed to evaluate, in some capacity, the potential for LLM systems to enable worst-case misuse risks by an attacker or threat actor. Most research teams included domain experts ($n = 12$) and half included social scientists ($n = 8$). All studies except one were randomized controlled trials with at least two arms (LLM access vs. control) ($n = 15$). Sample sizes in discussed uplift studies ranged from under 20 to nearly 5,000 participants (median = 110), with only three studies exceeding 1,000 participants. Studies recruited participants primarily

interviews (Hennink and Kaiser 2022; Guest, Bunce, and Johnson 2006). We observed that later interviews surfaced few new themes or challenges beyond those identified in earlier interviews, consistent with content saturation, though we did not conduct a formal saturation analysis. Our 30% participation rate was comparable to that of similar studies (Harrap et al. 2023).

Validity type	Definition and common threats
Construct [C]	The extent to which study operations (e.g., interventions, measures, settings, participants) correspond to intended abstract constructs (Campbell and Cook 1979; Cook, Campbell, and Shadish 2002; Salaudeen et al. 2025). Threats arise when the treatment is ill-defined (e.g., the LLM intervention updates mid-study), when control conditions fail to represent the relevant counterfactual, or when outcome measures capture only a subset of pathways relevant to the decision-relevant construct.
Internal [I]	The extent to which the assumptions required to identify a causal effect between the explanatory variable and the outcome of interest are satisfied within the study context (Imbens and Rubin 2015; Cook, Campbell, and Shadish 2002). Threats arise when treatment diffuses across experimental boundaries through spillovers or contamination, or when participants are differentially exposed to varying versions of an intervention.
External [E]	The extent to which internally valid causal effects generalize to different individuals, contexts, and outcomes (Meyer 1995; Cook, Campbell, and Shadish 2002). Threats arise when the recruited sample diverges from the decision-relevant population, when user proficiency co-evolves with the technology, or when baselines shift over time.

Table 2: Validity dimensions used to organize methodological challenges in human uplift studies.

Level	Definition
LLM-native ***	The causal mechanism generating the challenge is specific to properties of LLM-based systems; precedents in prior RCT contexts are weak or absent.
LLM-transformed **	Precedent exists in prior RCT contexts, but standard mitigations are inadequate for preserving the relevant form(s) of validity in LLM studies.
LLM-intensified *	Precedent exists in prior RCT contexts, and standard mitigations remain largely adequate; LLMs primarily increase the frequency or severity of the threat.

Table 3: Three-level classification of how distinctly each methodological challenge emerges from properties of LLM-based systems.

via convenience sampling through partner organizations, social media, or targeted outreach.

4.2 Thematic analysis

We present our thematic analysis in two stages. First, we present a structured synthesis of methodological challenges, organized across the AI evaluation lifecycle as defined in (Wei et al. 2025b; Paskov et al. 2025). We annotate challenges with [C], [I], or [E] to indicate the primary forms of validity they threaten, as defined in Table 2. We further tag challenges with a star rating (★ to ★★★) to indicate their degree of LLM-specificity, as defined in Table 3. Some challenges, like intervention fidelity (C5), are distinct to LLM-based RCTs. Others are *familiar in form but transformed* by properties of frontier AI systems, such as control conditions (C2) and interference (C6). Still others, such as interpretation of results across models and time (C9), are well-established in experimental design but take on distinct form in LLM uplift studies, particularly in high-stakes governance contexts. We then summarize a key subset of solutions. Bolding within quotations is added for emphasis. Table 1 maps challenges to corresponding solutions.

Methodological Challenges

C1 Measurement Under Constrained and Incomplete Pathway Coverage (Design) [C] ★ A growing body of evidence documents limitations of measurement instruments in AI evaluation (Raji et al. 2021; Hutchinson et al. 2022; Reuel et al. 2024; Rauh et al. 2024; Eriksson et al. 2025;

Wallach et al. 2025; American Statistical Association 2016), with human uplift studies as no exception. Experts described challenges in defining measurement instruments that proxy real-world behavior, particularly in safety and misuse-focused uplift studies, where action spaces are wide, settings are complex and adversarial, and motivating threat models are themselves often under-specified (Bengio et al. 2026). Several experts emphasized concerns about tasks’ experimental realism. Expert N discussed a cybersecurity study:

*N: “One big challenge we’re facing is whether or not the tasks, the actual lab environments that we’re giving these proxy attackers is realistic enough of the real world. Are we representing the real cyber world in assigning these tasks? That’s one big challenge. **And the implication is, if it’s not realistic, then who cares about the uplift that’s given?**”*

Experts noted that concerns about task realism are compounded by the need to constrain action spaces in order to make studies tractable. In domains with wide action spaces and multiple viable strategies, researchers must often focus on a subset of pathways, introducing trade-offs between granularity and coverage. Expert A remarked:

*A: “Studies need to artificially constrain [misuse scenario] pathways in order to be able to study anything at all...I think it’s important that we don’t lose sight of that. Say that you have specified one particular pathway and you didn’t see uplift in that, **that doesn’t need to necessarily mean that you’re now able to rule out***

Expert ID	Organization Type	Expert ID	Organization Type
Expert A	Independent Research Institution	Expert I	University
Expert B	University	Expert J	University
Expert C	Independent Research Institution	Expert K	University
Expert D	Other	Expert L	AI Company
Expert E	Independent Research Institution	Expert M	Independent Research Institution
Expert F	Independent Research Institution	Expert N	Independent Research Institution
Expert G	Independent Research Institution	Expert O	University
Expert H	Governmental Institution	Expert P	University

Table 4: Overview of experts by organization type.

all the other pathways that you didn’t study.”

Together, these challenges underscore that constrained and incomplete pathway coverage is often unavoidable in human uplift studies, but remains methodologically defensible only when task constraints are clearly specified, transparently communicated, and explicitly aligned with the research question of interest. This distinction is especially salient in governance and policy contexts, where results from a single constrained pathway may be interpreted, implicitly or explicitly, as speaking to a broader space of real-world behaviors or risks than that which the study was designed to measure.

C2 Control Conditions in AI-Integrated Environments (Design) [C] ★★ Selecting appropriate control conditions, defined here as the within-study comparison group used to identify the causal effect of AI systems, emerged as a central design challenge shaping both the interpretation and comparability of human uplift studies.⁶ While control specification is a general concern in randomized experiments, experts noted its unique challenge in human uplift studies, given the increasing embeddedness of AI tools in everyday workflows. In such settings, identifying a realistic and meaningful counterfactual is often nontrivial.

Experts described wide variation in control definitions in practice. Some studies restrict participants to basic internet search or non-AI tools, while others provide access to human experts or alternative software systems. Experts noted that, unlike many traditional interventions, LLMs do not replace a single prior tool but are layered onto an existing ecosystem of technologies. As such, control conditions are inherently relative to a reference point, which, if not made explicit, can lead to misinterpretation of uplift estimates. Expert P noted:

P: “In any control setting, there’s going to be some technology that they have available to them, whether it be AI, or it used to be called AI but now it’s no

longer AI, or maybe it’s some other AI tool...it’s always relative to something and maybe impacts interpretation. Think about the coding papers, for example, that study how Copilot impacts programmers. Well, before Copilot, there was Autocomplete and TabComplete and IDEs...”

The appropriate choice of control conditions ultimately depends on the specific research question and the reference point or threshold of interest, for which experts suggest scientific consensus or stakeholder consultation may be useful.

C3 Recruiting Populations Aligned with Specialized Research Questions (Recruitment) [E] ★ Experts described recruitment as a central constraint on external validity, particularly when research questions target specialized or high-stakes domains such as biosecurity, cybersecurity, or law. Experts described two recurring recruitment constraints. First, in many governance-, safety- and security-oriented studies, the populations of greatest interest are not directly recruitable at all. When research seeks to model malicious or highly capable threat actors, researchers must rely on proxy populations such as students, professionals, or domain experts, whose motivations, incentives, and constraints differ systematically from those of real-world adversaries. In these cases, uplift estimates depend on how well study designs, incentives, and task framing approximate relevant behaviors.

Second, in other settings, the population of interest is often difficult or costly to recruit. For example, Expert E examined AI use in legal contexts but recruited law students rather than practicing lawyers, noting that “lawyers’ time at law firms is pretty costly ... you just run up against budget constraints and have to make these trade-offs.” Similar challenges arise in domains requiring specialized technical expertise or narrow sub-disciplinary knowledge. Researchers often trade representativeness for feasibility, sometimes favoring more curated samples to reduce variance or improve retention at the cost of smaller or less representative populations. Across scenarios, recruitment choices shape results. In policy-relevant settings, where decisions frequently hinge on rare or adversarial behaviors rather than average effects, failure to clearly communicate recruitment limitations can lead to overconfident extrapolation about populations or behaviors.

C4 Heterogeneous and Evolving AI Literacy (Recruitment) [I, E] ★★ Participant proficiency in the use of AI, or *AI literacy*, emerged as a salient consideration in the design

⁶Throughout the paper, we distinguish *control conditions*, which are study-specific comparison groups used for within-study causal identification, from *baselines*, which are reference points used to contextualize and compare uplift magnitudes across studies or over time. We further refer to *human baselines* as baselines anchored to prevailing human capabilities or workflows (Wei et al. 2025b). We avoid using “baseline” to denote pre-intervention measurements, though this too is a common use case in literature.

and interpretation of human uplift studies. Experts emphasized that variation in AI literacy is especially consequential, given the degree to which performance depends on users' prompts, output interpretation, and tool integration. Expert L noted:

L: *"If you're selecting someone who is a complete novice, has never used GenAI, I think you will not see much uplift just because they don't know how to use the tool, whereas if you just give the same person six months to learn about GenAI and ask them again the same question six months down the road, they might be successful in using GenAI in a particular manner."*

Heterogeneity in AI literacy poses distinct threats to both external and internal validity. If the population of interest is mis-specified, such that participant skill does not reflect the population of interest, uplift estimates may fail to generalize, undermining external validity. If variation in AI literacy is unevenly distributed or not controlled for within a study, it may act as a confounder, threatening internal validity by obscuring the causal relationship between AI assistance and observed outcomes.

C5 Intervention Fidelity Under Rapid Model Evolution (Execution) [C, I] *** Experts identified intervention fidelity as a key challenge: unlike traditional interventions, the LLM under study can change mid-experiment, silently and asymmetrically (see, e.g., Kapoor et al. 2025; Zhu et al. 2026). In human uplift studies, the intervention typically includes access to a particular AI model, often embedded within a broader tool ecosystem including system prompts, safety filters, plugins, or auxiliary tools. Experts noted that developer or provider changes to any of these underlying components, while often unflagged, can materially affect study outcomes. A lack of clear visibility into model versioning, common among experts, induces uncertainty about intervention fidelity. Expert N noted:

N: *"When we started talking about this experiment and designing it, we noticed that the publicly available models were capable of running code in their environments and installing different Python tools. For example, we are starting [now] to see a lot more refusals from the same exact model. So the model has undergone an update and we no longer have access to that snapshot of the previous instance of that same model. And so if you run a study over a period of three months in which that model is being updated and you're unaware, you're comparing apples and oranges."*

When participants are exposed to materially different model versions or configurations at the same time, this introduces unbalanced heterogeneity in the intervention, threatening internal validity by violating the assumption that treatment is consistently defined across subjects. By contrast, when the intervention changes uniformly over time, for example, due to a globally deployed model update, internal validity may be preserved, provided that participants are affected symmetrically. However, rather than estimating the effect of a single, fixed model, the study would then capture

the effect of exposure to a changing system over a specified time horizon. The risk of such changes increases with study duration: brief laboratory tasks may face minimal exposure, while longitudinal or real-world studies spanning weeks or months are more vulnerable.

C6 Interference: Spillovers and Contamination in AI-Integrated Environments (Execution) [I] ** Experts highlighted challenges arising from spillovers and contamination in AI-integrated environments. *Spillovers* occur when exposure to AI diffuses indirectly, for example through social interaction, shared strategies, or collaboration, violating standard non-interference assumptions of RCTs. These risks are especially salient in settings with close cohort structures, such as classrooms, labs, workplaces, or training programs, where participants naturally exchange information. *Contamination*, by contrast, occurs when control group participants directly access restricted AI tools or comparable systems, violating experimental protocol. As Expert A noted:

A: *"I expect cheating to be much more salient in LLM uplift studies, especially if there is an internet-only control group. Contrast this with a clinical drug trial, where, if you're not giving the control group the drugs, they're probably not going to be able to acquire it."*

Across both dynamics exists a trade-off: study designs that aim to improve external validity by approximating real-world, longer-term use often face heightened risks of spillovers and contamination. Short, tightly controlled studies, on the other hand, may better preserve causal identification, but fail to reflect real-world AI use.

C7 Expectancy Effects (Execution) [I, E] ** Experts discussed limits to blinding in human uplift studies: because interaction with AI systems is explicit, conversational, and central to task execution, blinding participants to AI treatment is often infeasible. As a result, observed effects may reflect not only the technical capabilities of the system, but also users' expectations about what AI can do. In experimental psychology, such dynamics are often described as *expectancy effects* (e.g. (Klein et al. 2012)). Expectancy effects complicate interpretation when uplift results are used to make claims about underlying model performance or to support cross-study comparisons. Experts discussed partial mitigations – such as blinding outcome assessors or analysts – though such measures do not fully resolve these dynamics. Instead, the central interpretive choice lies in whether expectation-driven behavior is treated as a confound to be minimized and controlled for or as an integral feature of realistic AI use. Blinding challenges limit internal validity when uplift estimates are interpreted as isolating model capability, and external validity when results are compared across studies with different framing, populations, or deployment contexts.

C8 Documentation Under Proprietary and Security Constraints (Documentation) [E] * Experts described documentation — or the reporting of evaluation results, metrics, methodological details, experimental materials, and datasets — as a challenge to understanding evidence and

advancing methodologies. While evaluations in many fields follow established documentation standards (Hopewell et al. 2025; American Economic Association 2025), human uplift studies fall short due to their relative nascency, along with proprietary and security concerns. Expert D noted:

D: *“It’s often very difficult to know what was actually done on an uplift study based on model cards. Did they have a day to do the task? A week? Was it a multiple choice exam? An intensive planning exercise?”*

This lack of detail hampers scientific progress and appropriate interpretation of results for governance applications and beyond. While Expert B noted contractual solutions can ensure publishing rights, the feasibility of such arrangements can depend on power dynamics and dependencies between researchers, evaluators, and model developers, which shape the negotiability of terms in practice.

C9 Interpreting Results Across Models and Time (Documentation) [E] ★ While LLM uplift studies estimate a point-in-time effect of a specific model for a specific population on a specific task, their results often inform broader and longer-term governance questions on, for example, safeguards and deployment. The dynamics and constraints discussed in this section introduce challenges to understanding to what extent a given study’s estimates extend beyond its original conditions. Expert A noted:

A: *“There’s often a conflation of what study questions are of interest. There’s this overall question: how good are people at using AIs for this task right now? But often the study question that people want to answer in a safety context is: if we launch this model, and then in the next few years people become much better at using AIs, how well would they perform? ... I wish that studies could better isolate those two different questions and communicate clearly how their results should be interpreted.”*

The implications are greater when uplift results inform high-stakes decisions such as in safety cases and policy frameworks. Expert D illustrated how quickly the underlying conditions drift:

D: *“Comparison over time is very difficult...if you run another study with the same group of people, six months later, the world has probably changed in pretty meaningful ways. People are more familiar with using LLMs...That’s going to change the way that they perform.”*

Interpretive challenges further compound in cross-model comparisons. Unlike MCQA benchmarks such as MMLU (Hendrycks et al. 2021) or GPQA (Rein et al. 2023), human uplift studies cannot be easily rerun on updated models or alternative systems without substantial recruitment and execution costs. As a result, direct comparison across model versions is often infeasible in practice. Expert O explained:

O: *“If you ran a study with this model and then reviewers are like, well, why don’t you try model X or*

Y? It’s not a static benchmark that you can just re-run. You’d have to recruit a new group of participants, which is not necessarily realistic.”

Interpretive challenges similarly extend to baseline specification, where baseline refers to the reference point used to contextualize and compare uplift magnitudes across studies or over time. As decision-relevant comparison groups shift, driven by improvements in open-source models or the increasing integration of AI into everyday tools, reliance on outdated baselines risks comparisons that no longer reflect realistic conditions. Simultaneously, this dynamic risks what Expert D characterized as a “boiling frog” problem, in which gradual changes in reference points obscure substantial shifts in absolute capability (Kapoor et al. 2024).

Practical Solutions for the Field Experts proposed a range of solutions associated with the challenges above, spanning study-level practices to ecosystem-level interventions. We map these solutions to challenges below and in Table 1. In doing so, we aim to illustrate potential points of leverage rather than to advocate for any specific approach. This mapping is not exhaustive, prescriptive, or definitive.

S1 Standardized Task Libraries Experts proposed developing shared benchmark tasks for different task domains, given “huge economies of scale” and suggested that consortia of organizations could jointly contribute tasks for the common good without bearing full individual costs. Experts emphasized the importance of determining which tasks are representative or appropriately difficult. To this end, Expert A suggested consulting domain experts and surveying study designers, decision-makers, and relevant communities about proxy quality prior to study launch, noting that pre-registered expectations about real-world implications could help structure post-results interpretation. Expert B proposed using multiple KPIs to capture distinct dimensions of performance, while other experts similarly highlighted sub-task-level measurement as a way to preserve informative signal in complex, multi-step tasks or capabilities. More thoughtfully designed, crowd-sourced standardized task libraries can deepen and broaden measurement by encouraging systematic coverage of relevant questions and pathways and improving the validity of proxies through economies of scale. **Related challenges: C1, C8, C9**

S2 Baseline and Control Selection Conventions Experts advocated for clearer conventions distinguishing baseline and control selection to improve interpretability and comparability across uplift studies. Suggested practices included explicitly characterizing the *prior technology or workflow* being displaced when defining design and, where feasible, adopting standardized baseline and control bundles for common settings (e.g., academic research, professional services). Experts emphasized that both baseline and control choices should be made explicit and align with the specific reference point or decision threshold relevant to the downstream use of the results. Explicit baseline and control conventions could anchor results to a clear reference point, improving interpretability and cross-study comparability. Further, conventions could align evaluations with

decision-relevant questions, support consistent documentation through shared standards, and foster crowd-sourced identification and mitigation of spillovers or contamination. **Related challenges:** C2, C6, C8, C9

S3 Leveling and Accounting for AI Literacy Experts reported a range of strategies for AI literacy challenges, including measuring and filtering for proficiency in recruitment, stratifying randomization on prior experience, providing introductory training to participants, and controlling for AI skill level in post-hoc analysis. Accounting for AI literacy makes user heterogeneity explicit, helping align research questions with the populations and interaction modes the study seeks to capture, while potentially mitigating effects of confounding variables. Clear documentation of skill supports transparent reporting and fosters more appropriate interpretation by clarifying the user groups to which results apply. **Related challenges:** C4, C8, C9

S4 Versioned Evaluation Infrastructure, Snapshot Access, and Verification Versioned AI systems and stable evaluation snapshots could allow researchers to measure, control for, or appropriately interpret intervention fidelity. Provider guarantees of fixed model versions for study duration, explicit version identifiers for model/system configurations, and mediated access to snapshots through secure or tiered research environments could improve research rigor (Casper et al. 2024; Brundage et al. 2026). The success of this solution hinges on coordination between developers, providers, and researchers. Such infrastructure need not imply public release of sensitive systems, but could instead support intervention fidelity, reproducibility, and interpretability under proprietary or security constraints. Emerging research on verification offers potential solutions (Balan, Learney, and Wood 2025; Scaramuzza et al. 2026; Scaramuzza, Quattrocchi, and Tamburri 2025; Yao and Juarez 2026; You 2026). **Related challenges:** C5, C8, C9

S5 Contamination and Spillover Management Experts proposed mitigating contamination and spillover threats through multiple approaches: monitoring LLM usage by control groups and removing protocol violators; offering post-study “amnesty” periods where participants can admit to violations without penalty while allowing data exclusion; physically separating treatment and control groups through staggered scheduling or dispersed recruitment; and implementing technical controls such as providing restricted Chromebooks or network-level site blocking. Several experts emphasized the importance of designing incentive structures that encourage compliance rather than relying solely on monitoring and detection systems.⁷ Contamination and spillover management practices can mitigate interference issues and promote better documentation and interpretation of results. **Related challenges:** C6, C8, C9

S6 Natural Experiments Expert B identified phased product roll-outs and staggered deployments as underuti-

⁷Notably, in contexts where subjects lack strong performance incentives, contamination may be less problematic. This highlights a fundamental trade-off: while incentives simulate real-world motivations, they may also incentivize control groups to cheat.

lized opportunities for generating quasi-experimental variation in human uplift studies, noting “it’s not that expensive because [companies] are going to [deploy in phases] in many cases — you can’t give [the new technology] to everybody all at once anyways. So if they coordinate with academics, I think there’s a big opportunity to ... get experimental variation.” When roll-out timing or access is determined by operational constraints such as infrastructure limits, geographic sequencing, or user tiering, rather than individual characteristics, these settings can approximate natural experiments that support causal inference under weaker assumptions than fully randomized trials. In such cases, differences in exposure may plausibly be treated as exogenous to user ability or motivation, enabling estimation of uplift effects with reduced selection bias. Natural experiments constrain the research question to real-world adoption contexts and can sidestep recruitment challenges, given built-in inclusion of decision-relevant populations. Natural experiments can also support interpretation of results across models or deployment stages. **Related challenges:** C3, C9

S7 AI-Accelerated Research Methods Experts identified AI as a promising tool for addressing scalability challenges in uplift research, highlighting its potential to reduce timeline and cost constraints. For study design, Expert L suggested using AI agents as complementary participants in pilot studies to enable rapid testing of experimental protocols (e.g., Ruan et al. 2024; Anthis et al. 2025; Lu et al. 2025; Froger et al. 2025; Park et al. 2026). Agent-based pilots could help, e.g., surface brittle task designs, unintended affordances, or reward-hacking in complex multi-step settings at low cost. AI-accelerated methods may also enable rapid exploration of questions and behavior across diverse simulated human profiles under clean experimental conditions. At the same time, agent behavior may be limited in external validity and should be cautiously vetted, validated, and used to complement rather than substitute for human subjects research. **Related challenges:** C3, C4, C5, C6, C9

S8 Post-Hoc Analysis Post-hoc analytical techniques can partially mitigate internal validity challenges in human uplift studies, particularly when ideal experimental control is infeasible. Careful ex post analysis, such as deriving heterogeneous treatment effects, adjusting for non-compliance, or reweighting observations to account for imbalance or attrition, can help diagnose and, in some cases, adjust for violations of key identification assumptions. Several experts emphasized that uplift studies generate unusually rich interaction logs, which enable finer-grained post-hoc audits of participant behavior. Expert A further noted that adhering to research standards, such as explicitly estimating causal effects rather than relying solely on directional hypothesis tests, can strengthen internal validity and downstream interpretation. While post-hoc analysis cannot substitute for robust experimental design, experts viewed it as a valuable complement. **Related challenges:** C3, C4, C5, C6, C7, C9

S9 Information Security Advisory Frameworks and Tiered-Access For studies presenting security or proprietary concerns, Expert A advocated for structured, deliber-

ate consultation with relevant communities through advisory boards that include domain and scientific experts. For example, human uplift studies on medical tasks may draw from consensus among medical experts, while human uplift studies on bioterrorism tasks may draw from consensus among national security advisors. This approach could help balance security, proprietary, and transparency concerns by providing independent review of information hazard risks while maintaining scientific openness where appropriate. Experts framed documentation not as a binary choice between openness and secrecy, but rather as a question of granularity and audience. Tiered-access reporting provides a practical mechanism for disclosing sufficient information to support interpretation of uplift results, while responsibly managing proprietary and security risks through controlled access (e.g., Frontier Model Forum 2025a). **Related challenges: C8, C9**

5 Discussion: Validity Trade-offs, Collective Action, and Governance

Expert interviews surfaced three recurring methodological tensions in human uplift research. Each implicates construct [C], internal [I], and external validity [E], though through different mechanisms. *Realism versus control*, for example, trades ecological realism against causal identification: naturalistic settings invite the spillovers (C6) that threaten causal identification, while the controlled tasks that protect it may poorly reflect real use (C1). *Speed versus rigor* trades timely evidence against design depth: rigorous, externally valid studies of long-horizon tasks may require months or years from design to completion, a window during which the intervention and research question itself may change (C5, C9). *Disclosure versus protection* affects the assessment of all three validity types, with proprietary and security constraints (C8) limiting the reporting and documentation needed to effectively interpret and act on evidence.

These tensions carry direct consequences for AI governance: studies designed on weak foundations or reported without sufficient context may produce false confidence or, conversely, trigger unnecessarily burdensome obligations. Well-designed and clearly documented studies, by contrast, can strengthen the evidence base for governance. Crucially, validity concerns do not operate in isolation: design choices that strengthen one form of validity often weaken another. Policy decisions should therefore rest on convergent evidence from multiple studies using complementary approaches. In governance decisions in which the relevant question shifts faster than any one study can resolve it, uplift estimates may be combined with approaches such as scenario analysis, forecasting, and methods designed to inform decisions under deep uncertainty (Lempert 2025; Marchau et al. 2019). In all cases, rigorous documentation to address C9, paired with tiered-access reporting (S9) where relevant, is a precondition for effective use of uplift evidence.

While some challenges (e.g. C1, C3, C8, C9), familiar in form, can turn to established methodological precedent for some guidance (e.g., Meyer 1995; Hopewell et al. 2025; Hoffmann et al. 2014), a novel set of challenges stems from LLM-specific properties (e.g. C5 and, to varying degrees,

C2, C4, C6, C7). For these challenges, existing methodological tools offer weaker guidance. The solutions proposed by interviewed experts represent early attempts to fill this gap. Each solution carries its own validity trade-offs, which may be empirically explored in future work.

Beyond any individual uplift study lies a meta-level need for collective action, on which several proposed solutions (S1, S2, S9) depend. Without coordination, challenges C1–C9 are more likely to persist. At present, fragmented infrastructure forces research teams to reinvent methods, while proprietary evaluations and misaligned research incentives constrain knowledge-sharing. Field-level workshops, shared infrastructure, and consensus-building processes could enable researchers to co-develop novel solutions for AI evaluation and governance. Fields with deep experience in RCTs and causal inference (e.g. medicine, psychology, and economics) remain severely underleveraged within human uplift research: cross-domain synergies and bridges could meaningfully advance human uplift, evaluation, and AI governance. Here, public and philanthropic actors are uniquely positioned to support evaluations and initiatives that no single organization has incentives to undertake.

6 Conclusion

Human uplift studies, or studies that measure AI effects on human performance relative to a counterfactual, offer a framework for measuring the societal impacts of AI systems. By studying how AI impacts human performance in realistic tasks, uplift studies can inform governance decisions in ways that model-centric static benchmarks cannot. At the same time, the application of RCT methodologies to frontier AI systems introduces distinctive challenges that complicate both study design and interpretation and, subsequently, threaten to confound downstream decision-making.

Across 16 expert interviews, we find that these challenges map onto familiar threats to construct, internal, and external validity, amplified by the properties of frontier AI systems. These trade-offs imply that no single uplift study, however well designed, can provide definitive assessments of system safety or societal impact. Policy-relevant conclusions should thus rest on convergent evidence drawn from multiple studies with complementary designs, not on isolated findings.

Looking forward, advancing the rigor and usefulness of human uplift research will require more than incremental improvements in individual study design. Established disciplines hold relevant expertise and experts identified a range of novel and practical solutions, including standardized task libraries, clearer baseline and control conventions, versioned evaluation infrastructure, and structured approaches to interference, that can partially mitigate these challenges. Most importantly, the field requires coordination mechanisms to address collective action problems, enabling knowledge and resources to accumulate rather than remain siloed.

As human uplift studies play a growing role in informing AI policy, strengthening the methodological foundations of uplift research and clarifying what these studies can and cannot support becomes essential for responsible governance. Doing so can contribute to a coherent and policy-relevant body of evidence towards more beneficial AI futures.

Acknowledgements

The authors are grateful to the expert interviewees who generously volunteered their time, experience, and judgment to participate in this study. Their insights substantially informed both the framing of the research and the interpretation of its findings. The authors also thank the leadership of RAND's Center on AI, Security, and Technology (CAST) for their support of this work. RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken through CAST, which examines the opportunities and risks of rapid technological change, with a focus on artificial intelligence, security, and biotechnology. For more information, contact cast@rand.org.

Ethical Considerations Statement

This study was reviewed by RAND's Institutional Review Board and deemed exempt under 2024-N0632-MOD-06. The research draws on semi-structured interviews with expert practitioners and involved no experimentation, intervention, or deception. All experts provided informed consent prior to participation and were informed of the voluntary nature of the study, their right to decline to answer questions, and their ability to withdraw at any time without penalty. Interviews were conducted using secure videoconferencing infrastructure and, with expert permission, audio-recorded for transcription. Raw recordings and non-anonymized transcripts were accessible only to authorized project staff and stored on encrypted organizational systems. Recordings were deleted following transcription and anonymization. Anonymized transcripts were reviewed to remove identifying details prior to analysis or sharing with collaborators. Experts were asked how they wished to be identified in any published materials, and no quotations were attributed to individuals by name. All quotations included in the paper were cleared with experts. These procedures were designed to minimize risks of inadvertent disclosure or reputational harm while enabling transparent reporting of methodological insights.

Funding

This research was independently initiated and conducted within the RAND's Center on AI, Security, and Technology (CAST) using income from operations and gifts and grants from philanthropic supporters. A complete list of donors and funders is available at www.rand.org/CAST. RAND clients, donors, and grantors have no influence over research findings or recommendations.

References

Agrawal, N.; Binns, R.; Van Kleek, M.; Laine, K.; and Shadbolt, N. 2021. Exploring Design and Governance Challenges in the Development of Privacy-Preserving Computation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8096-6.

American Economic Association. 2025. About the AEA RCT Registry.

American Statistical Association. 2016. American Statistical Association Releases Statement on Statistical Significance and P-Values.

Anderson-Samways, B.; Ee, S.; O'Brien, J.; Buhl, M.; and Williams, Z. 2024. Responsible Scaling: Comparing Government Guidance and Company Policy.

Angrist, J. D.; and Pischke, J.-S. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. ISBN 978-0-691-12035-5.

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Evans, J.; Brynjolfsson, E.; and Bernstein, M. 2025. LLM Social Simulations Are a Promising Research Method. arXiv:2504.02234.

Anthropic. 2023. Anthropic's Responsible Scaling Policy.

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4.

Balan, K.; Learney, R.; and Wood, T. 2025. A Framework for Cryptographic Verifiability of End-to-End AI Pipelines. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, IWSPA '25, 49–59. New York, NY, USA: Association for Computing Machinery. ISBN 9798400715013.

Banerjee, A. V. 2020. Field Experiments and the Practice of Economics. *American Economic Review*, 110(7): 1937–1951.

Banerjee, A. V.; Duflo, E.; and Kremer, M. 2020. The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. In Basu, K.; Rosenblatt, D.; and Sepúlveda, C., eds., *The State of Economics, the State of the World*, 439–487. The MIT Press. ISBN 978-0-262-35347-2.

Bassen, J.; Balaji, B.; Schaarschmidt, M.; Thille, C.; Painter, J.; Zimmaro, D.; Games, A.; Fast, E.; and Mitchell, J. C. 2020. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.

Becker, J.; Rush, N.; Barnes, E.; and Rein, D. 2025. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. ArXiv:2507.09089 [cs].

Bengio, Y.; Clare, S.; Prunkl, C.; Andriushchenko, M.; Bucknall, B.; Murray, M.; et al. 2026. International AI Safety Report 2026. Technical Report DSIT 2026/001, UK Department for Science, Innovation and Technology. Accessed: 2026-05-20.

Bengio, Y.; Fox, B.; Carlos Ponce de Leon Ferreira de Carvalho, A.; Nemer, M.; Rivera, R. P.; Zeng, Y.; Heikkilä, J.; Avrin, G.; Krüger, A.; Ravindran, B.; Riza, H.; Seoighe, C.; Katzir, Z.; Monti, A.; Kitano, H.; Kerema, M.; López Portillo, J. R.; Sheikh, H.; Jolly, G.; Ajala, O.; Ligtot, D.; Lee, K. M.; Hatip, A. H.; Rugege, C.; Albalawi, F.; Wong, D.; Oliver, N.; Busch, C.; Molchanovskiy, O.; Alserkal, M.;

- Khan, S. M.; McLean, A.; Gill, A.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Goldfarb, D.; Heidari, H.; Khalatbari, L.; Longpre, S.; Mavroudis, V.; Mazeika, M.; Ng, K. Y.; Okolo, C. T.; Raji, D.; Skeadas, T.; Tramèr, F.; Adekanmbi, B.; Christiano, P.; Dalrymple, D.; Dietterich, T. G.; Felten, E.; Fung, P.; Gourinchas, P.-O.; Jennings, N.; Krause, A.; Liang, P.; Ludermir, T.; Marda, V.; Margetts, H.; McDermid, J. A.; Narayanan, A.; Nelson, A.; Oh, A.; Ramchurn, G.; Russell, S.; Schaake, M.; Song, D.; Soto, A.; Tiedrich, L.; Varoquaux, G.; Yao, A.; Zhang, Y.-Q.; and Prud'homme, B. 2024. International Scientific Report on the Safety of Advanced AI: Interim Report. Research Report DSIT 2024/009, AI Seoul Summit, Seoul, Republic of Korea.
- Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101. Publisher: Routledge .eprint: <https://doi.org/10.1191/1478088706qp063oa>.
- Braun, V.; and Clarke, V. 2023. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health*, 24(1): 1–6. Publisher: Taylor & Francis .eprint: <https://doi.org/10.1080/26895269.2022.2129597>.
- Braun, V.; Clarke, V.; Hayfield, N.; and Terry, G. 2024. Thematic Analysis. In *Advanced Research Methods for Applied Psychology*. Routledge, 2 edition. ISBN 978-1-003-36271-5. Num Pages: 11.
- Brodeur, A.; Valenta, D.; Marcoci, A.; Aparicio, J. P.; Mikola, D.; Barbarioli, B.; Alexander, R.; Deer, L.; Stafford, T.; Vilhuber, L.; Bensch, G.; Goldschmitt, D.; Gourdon-Kanhukamwe, A.; de Varda, A. G.; Grigoryeva, I.; Gugushvili, A.; Fletcher, A. H. A.; Habermann, F.; Hablicsek, M.; Haddad, J.; Hall, J. D.; Abdelhady, M.; Hammar, O.; Hassouneh, M.; Hausladen, C. I.; Hendrikse, S. C. F.; Hepplewhite, M.; Ho, A. T. Y.; Hogan-Hennessy, S.; Howley, E.; Huang, G.; Hulstaert, H.; Abdelmoula, Y.; Ilchovska, Z. G.; Jakobsson, N.; Jansson, J.; Jarosz, E.; Jebeli, H.; Jiang, Y.; Junaid, H.; Kalluraya, R.; Kelly, E.; Kimel, E.; Abdul Baki, G.; Kingsuwanikul, S.; Klotzbücher, V.; Krähmer, D.; Krüminas, P.; Kruus, N.; Kujansuu, E.; Kurz, C. F.; Küster, S.; Lee-Whiting, B.; Lewandowski, F.; Aguirre, T.; Li, T.; Li, R.; Liu, D.; Liu, J.; Lo, H.; Loter, K.; Dias, F. M.; Madan, C. R.; Mäder, N.; Mandas, M.; Aiyer, S.; Marcus, J.; Fages, D. M.; Martin, X.; McWay, R.; Medina-Gaspar, D.; Meng, S.; Meng, L.; Miller, A. P.; Mirabel, T.; Mishra, D. D.; Akhtar, S.; Mishra, S.; Moges, B. W.; Mojarrad, M. M.; Mohnen, M.; Morin, L.-P.; Motoki, F.; Muehlenbachs, L.; Mullin, G.; Musulan, A.; Muzzi, S.; Akhtar, F.; Myers, J. A. C.; Neubauer, F.; Nguyen, T.; Ni-azi, A.; Nordstrom, A.; Nowak, B.; O'Habib, D.; Ölkens, T.; Ong, J.; Orozco Castiblanco, V.; Albada, M. R.; Özak, Ö.; Ozkes, A. I.; Paaso, M.; Pandey, S.; Papazoglou, V.; Penheiro, R.; Pham, L.; Phielers, U.; Pütz, P.; Qi, Q.; Altman, M.; Qiu, J.; Reinstein, D. A.; Repo, J.; Rudolf, N.; Saha, S.; Saka, O.; Saponaro, C.; Sator, G.; Schoenmakers, M.; Seri, R.; Angenendt, D.; Shah, M.; Sibille, P.; Siemroth, C.; Skavysh, V.; Slater, B.; Song, W.; Staubli, S.; Steindl, T.; Waongo, N. S.; Stott, P.; Lari, Z. A.; Strobel, S.; Sudharan, R.; Sun, P.; Swain, S. D.; Talavera, O.; Tantiangco, H. M.; Tarasenko, G.; Tarlinton, B.; Tarraf, M.; Teoh, K.; De León Tejada, J. A.; Thériault, R.; Thompson, B.; Tian, T.; Tian, W.; Rein, M. T.; Tolani, E.; Topstad Borgen, N.; Topstad Borgen, S.; Torralba, J.; Velez-Ospina, C.; Asanov, I.; Mak, M. W.; Wallrich, L.; Wang, Z.; Ward, L.; Webb, M. D.; Webb, D.; Weber, B. S.; Weber, C.; Weng, W.-C.; Westheide, C.; Asanov Noha, A.-M.; Wilkinson, T.; Wong, K.-Y.; Wroński, M.; Wu, Z.; Wu, Q.; Wu, V. Y.; Xiao, B.; Xu, F.; Xu, C.; Yadav, P.; Ashong, R.; Chou, Y. Y.; Yap, L.; Yazbeck, M.; Zagrodzka, Z.; Zahra, T.; Zaneva, M.; Zhang, X.; Zhao, Z.; Zhong, H.; Zirculis, A.; Auer, T.; Zou, J.; Zoutman, F.; Zozoungbo, C.; Bahamonde-Birke, F. J.; Baker, B. J.; Bartram, S. M.; Bao, D.; Batinovic, L.; Batistoni, T.; Beeder, M.; Beland, L.-P.; Bienz, C.; Aryanto, C. B.; Bolibaugh, C.; Bonander, C.; Bravo, R.; Brennan, K.; Bronnikov, E.; Bruns, S.; Buliskeria, N.; Caicedo-Silva, S.; Calef, A.; Caulker, S.; Cepenas, S.; Chatton, A.; Chen, Z.; Ewurum, N. C.; Ciocîrlan, A.-B.; Clouth, F. J.; Collins, J.; Cook, N.; Cornejo, C.; Craveiro, J.; Cui, J.; Vayalabron, N. C.; Czymara, C.; Bermúdez Jaramillo, C. D.; Datta, H.; Denoo, L.; Dhaliwal, A.; Dhameja, N.; Djemai, E.; Dujeancourt, E.; Dündar, U.; Duprey, T.; Eissa, Y.; El Fassi, Y.; El Fassi, I.; Ellis, K.; Elminejad, A.; Elsherif, M.; Emirmahmutoglu, A.; Etingin-Frati, G.; Eze, E.; Dollbaum, J. F.; Victoria, J. F.; Rengifo Jaramillo, A. F.; Fenig, G.; Fernandes, V.; Fiala, L.; Fink, L.; Fish, S.; Fitzgerald, J.; Forshaw, R.; Fortier-Chouinard, A.; Fréget, L.; Frese, J.; Gabani, J.; Gallegos, S.; Gamill, M. C.; Gáspár, A.; Gauriot, R.; Gavrilova, E.; Gerales, D.; Cantone, G. G.; and Gibson, G. 2025. Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science. Working Paper 195, I4R Discussion Paper Series.
- Brooks, M. 2025. Is your AI benchmark lying to you? *Nature*, 644(8075): 294–296. Bandiera_abtest: a Cg_type: Technology Feature Publisher: Nature Publishing Group Subject_term: Technology, Machine learning, Mathematics and computing, Genomics.
- Brundage, M.; Dreksler, N.; Homewood, A.; McGregor, S.; Paskov, P.; Stosz, C.; Sastry, G.; Cooper, A. F.; Balston, G.; Adler, S.; Casper, S.; Anderljung, M.; Werner, G.; Mindermann, S.; Mavroudis, V.; Bucknall, B.; Stix, C.; Freund, J.; Pacchiardi, L.; Hernández-Orallo, J.; Pistillo, M.; Chen, M.; Painter, C.; Ball, D. W.; O'Keefe, C.; Weil, G.; Harack, B.; Finley, G.; Hassan, R.; Emmons, S.; Foster, C.; Reuel, A.; Treece, B.; Bengio, Y.; Reti, D.; Bommasani, R.; Trout, C.; Dattani, R.; Shamsabadi, A. S.; Weller, A.; Trager, R.; Tovcimak, R.; Papadatos, H.; and Murray, M. 2026. Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies. Working paper.
- Brynjolfsson, E.; Li, D.; and Raymond, L. 2025. Generative AI at work. *The Quarterly Journal of Economics*, 140(2): 889–942.
- Bucknall, B.; Siddiqui, S.; Thurnherr, L.; McGurk, C.; Harack, B.; Reuel, A.; Paskov, P.; Mahoney, C.; Mindermann, S.; Singer, S.; Hiremath, V.; Segerie, C.-R.; Delaney, O.; Abate, A.; Barez, F.; Cohen, M. K.; Torr, P.; Huszár,

- F.; Calinescu, A.; Jones, G. D.; and Trager, R. 2025. In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate? In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 3148–3161. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1482-5.
- Burden, J.; Tešić, M.; Pacchiardi, L.; and Hernández-Orallo, J. 2025. Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture. ArXiv:2502.15620 [cs].
- Buruk, O. O.; Matjeka, L. P.; and Mueller, F. F. 2023. Towards Designing Playful Bodily Extensions: Learning from Expert Interviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–20. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.
- California State Legislature. 2025. Senate Bill 53: Artificial Intelligence Models: Large Developers. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=20250260SB53. Accessed: 2026-03-10.
- Campbell, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4): 297–312.
- Campbell, D. T.; and Cook, T. D. 1979. Quasi-experimentation. *Chicago, IL: Rand McNally*, 1(1): 1–384.
- Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; Sharkey, L.; Krishna, S.; Von Hagen, M.; Alberti, S.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; Krueger, D.; and Hadfield-Menell, D. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2254–2272. ACM.
- Chen, V.; Talwalkar, A.; Brennan, R.; and Neubig, G. 2025. Code with Me or for Me? How Increasing AI Automation Transforms Developer Workflows. ArXiv:2507.08149 [cs].
- Cheong, I.; Xia, K.; Feng, K. J. K.; Chen, Q. Z.; and Zhang, A. X. 2024. (A)I Am Not a Lawyer, But...: Engaging Legal Experts towards Responsible LLM Policies for Legal Advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2454–2469. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0450-5.
- Choi, J. H.; and Schwarcz, D. 2024. AI Assistance in Legal Analysis: An Empirical Study. *Journal of Legal Education*, 73: 384.
- Chouldechova, A.; Atalla, C.; Barocas, S.; Cooper, A. F.; Corvi, E.; Dow, P. A.; Garcia-Gathright, J.; Pangakis, N.; Reed, S.; Sheng, E.; Vann, D.; Vogel, M.; Washington, H.; and Wallach, H. 2024. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. arXiv:2412.01934.
- Cook, T. D.; Campbell, D. T.; and Shadish, W. 2002. *Experimental and quasi-experimental designs for generalized causal inference*, volume 1195. Houghton Mifflin Boston, MA.
- Dahlinger, A.; Wortmann, F.; Ryder, B.; and Gahr, B. 2018. The Impact of Abstract vs. Concrete Feedback Design on Behavior Insights from a Large Eco-Driving Field Experiment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5620-6.
- Delaney, O.; Guest, O.; and Williams, Z. 2024. Mapping Technical Safety Research at AI Companies: A literature review and incentives analysis. arXiv:2409.07878.
- Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 473–484. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Dobbe, R.; Gilbert, T. K.; and Mintz, Y. 2021. Hard Choices in Artificial Intelligence. arXiv:2106.11022.
- Eble, A.; Boone, P.; and Elbourne, D. 2017. On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics.
- Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. arXiv:2502.06559.
- European Commission. 2025. Code of Practice for General-Purpose AI Models: Safety and Security Chapter.
- European Parliament; and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
- Farrell, H.; Gopnik, A.; Shalizi, C.; and Evans, J. 2025. Large AI models are cultural and social technologies. *Science*, 387(6739): 1153–1156.
- Feffer, M.; Sinha, A.; Deng, W. H.; Lipton, Z. C.; and Heidari, H. 2025. Red-Teaming for Generative AI: Silver Bullet or Security Theater? In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 421–437. AAAI Press.
- Friedler, S.; Singh, R.; Bili-Hamelin, B.; Metcalf, J.; and Chen, B. J. 2023. AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability.
- Froger, R.; Andrews, P.; Bettini, M.; Budhiraja, A.; Cabral, R. S.; Do, V.; Garreau, E.; Gaya, J.-B.; Laurençon, H.; Lecanu, M.; Malkan, K.; Mekala, D.; Ménard, P.; Bertran, G. M.-T.; Piterbarg, U.; Plekhanov, M.; Rita, M.; Rusakov, A.; Vorotilov, V.; Wang, M.; Yu, I.; Benhalloum, A.; Mialon, G.; and Scialom, T. 2025. ARE: Scaling Up Agent Environments and Evaluations. arXiv:2509.17158.
- Frontier Model Forum. 2025a. FMF Announces First-Of-Its-Kind Information-Sharing Agreement.
- Frontier Model Forum. 2025b. Frontier Capability Assessments. Technical report, Frontier Model Forum.
- Future of Life Institute. 2025. 2025 AI Safety Index. Technical report, Future of Life Institute.

- G. Mitchell, E.; M. Heitkemper, E.; Burgermaster, M.; E. Levine, M.; Miao, Y.; L. Hwang, M.; M. Desai, P.; Cas-sells, A.; N. Tobin, J.; G. Tabak, E.; J. Albers, D.; M. Smal-done, A.; and Mamykina, L. 2021. From Reflection to Ac-tion: Combining Machine Learning with Expert Knowledge for Nutrition Goal Recommendations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Sys-tems*, CHI '21, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8096-6.
- Google DeepMind. 2024. Frontier Safety Framework. <https://deepmind.google/blog/introducing-the-frontier-safety-framework/>. Accessed: 2026-05-20.
- Groves, L.; Metcalf, J.; Kennedy, A.; Vecchione, B.; and Strait, A. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1107–1120. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Guest, G.; Bunce, A.; and Johnson, L. 2006. How Many In-terviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*, 18(1): 59–82. Publisher: SAGE Publications Inc.
- Harrap, B.; Taylor, T.; Russell, G.; and Scott, A. 2023. A randomised controlled trial of email versus mailed invitation letter in a national longitudinal survey of physicians. *PLOS ONE*, 18(8): e0289628.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Mul-titask Language Understanding. arXiv:2009.03300.
- Hennink, M.; and Kaiser, B. N. 2022. Sample sizes for satu-ration in qualitative research: A systematic review of empir-ical tests. *Social Science & Medicine*, 292: 114523.
- Ho, A.; and Berg, A. 2025. Do the biorisk evaluations of AI labs actually measure the risk of developing bioweapons?
- Ho, A.; and Denain, J. S. 2025. The real reason AI bench-marks haven't reflected economic impacts.
- Hoffmann, T. C.; Glasziou, P. P.; Boutron, I.; Milne, R.; Per-era, R.; Moher, D.; Altman, D. G.; Barbour, V.; Macdonald, H.; Johnston, M.; Lamb, S. E.; Dixon-Woods, M.; McCul-loch, P.; Wyatt, J. C.; Chan, A.-W.; and Michie, S. 2014. Bet-ter reporting of interventions: template for intervention de-scription and replication (TIDieR) checklist and guide. *BMJ*, 348: g1687.
- Hopewell, S.; Chan, A.-W.; Collins, G. S.; Hróbjartsson, A.; Moher, D.; Schulz, K. F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J. A.; Bhandari, N.; Butcher, N. J.; Campbell, M. K.; Chidebe, R. C. W.; Elbourne, D.; Farmer, A.; Fer-gusson, D. A.; Golub, R. M.; Goodman, S. N.; Hoffmann, T. C.; Ioannidis, J. P. A.; Kahan, B. C.; Knowles, R. L.; Lamb, S. E.; Lewis, S.; Loder, E.; Offringa, M.; Ravaud, P.; Richards, D. P.; Rockhold, F. W.; Schrager, D. L.; Siegfried, N. L.; Staniszewska, S.; Taylor, R. S.; Thabane, L.; Torger-son, D.; Vohra, S.; White, I. R.; and Boutron, I. 2025. CON-SORT 2025 statement: updated guideline for reporting ran-domized trials. *Nature Medicine*, 31(6): 1776–1783. Pub-lisher: Nature Publishing Group.
- Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation Gaps in Machine Learn-ing Practice. arXiv:2205.05256.
- Ibrahim, L.; Huang, S.; Bhatt, U.; Ahmad, L.; and An-derljung, M. 2025. Towards interactive evaluations for in-teraction harms in human-AI systems. ArXiv:2405.10632 [cs].
- Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
- Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naa-man, M. 2019. AI-Mediated Communication: How the Per-ception that Profile Text was Written by AI Affects Trust-worthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–13. New York, NY, USA: Association for Computing Machin-ery. ISBN 978-1-4503-5970-2.
- Janjeva, A.; Gausen, A.; Mercer, S.; and Sippy, T. 2024. Evaluating Malicious Generative AI Capabilities. Technical report, The Alan Turing Institute.
- Jeong, S.; Breazeal, C.; Logan, D.; and Weinstock, P. 2018. Huggable: The Impact of Embodiment on Promoting Socio-emotional Interactions for Young Pediatric Inpatients. In *Proceedings of the 2018 CHI Conference on Human Fac-tors in Computing Systems*, CHI '18, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5620-6.
- Jones, C. R.; Rath, I.; Taylor, S.; and Bergen, B. K. 2025. People cannot distinguish GPT-4 from a human in a Turing test. In *Proceedings of the 2025 ACM Conference on Fair-ness, Accountability, and Transparency*, FAccT '25, 1615–1639. New York, NY, USA: Association for Computing Ma-chinery. ISBN 979-8-4007-1482-5.
- Kapoor, S.; Bommasani, R.; Klyman, K.; Longpre, S.; Ra-maswami, A.; Cihon, P.; Hopkins, A.; Bankston, K.; Bider-man, S.; Bogen, M.; Chowdhury, R.; Engler, A.; Henderson, P.; Jernite, Y.; Lazar, S.; Maffulli, S.; Nelson, A.; Pineau, J.; Skowron, A.; Song, D.; Storch, V.; Zhang, D.; Ho, D. E.; Liang, P.; and Narayanan, A. 2024. On the Societal Impact of Open Foundation Models. ArXiv:2403.07918 [cs] ver-sion: 1.
- Kapoor, S.; Stroebl, B.; Kirgis, P.; Nadgir, N.; Siegel, Z. S.; Wei, B.; Xue, T.; Chen, Z.; Chen, F.; Utpala, S.; Ndzomga, F.; Oruganty, D.; Luskin, S.; Liu, K.; Yu, B.; Arora, A.; Hahm, D.; Trivedi, H.; Sun, H.; Lee, J.; Jin, T.; Mai, Y.; Zhou, Y.; Zhu, Y.; Bommasani, R.; Kang, D.; Song, D.; Hen-derson, P.; Su, Y.; Liang, P.; and Narayanan, A. 2025. Holis-tic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. arXiv:2510.11977.
- Karnofsky, H. 2024. If-Then Commitments for AI Risk Re-duction. Technical report, Carnegie Endowment for Interna-tional Peace.
- Klein, O.; Doyen, S.; Leys, C.; de Saldanha da Gama, P. A. M.; Miller, S.; Questienne, L.; and Cleeremans, A. 2012. Low Hopes, High Expectations: Expectancy Effects and the Replicability of Behavioral Experiments. *Perspectives on Psychological Science*, 7(6): 572–584. PMID: 26168114.

- Krishna, S.; Mehrabi, N.; Mohanty, A.; Memelli, M.; Ponzo, V.; Motwani, P.; and Gupta, R. 2025. Evaluating the Critical Risks of Amazon's Nova Premier under the Frontier Model Safety Framework.
- Lee, H.; Stinar, F.; Zong, R.; Valdiviejas, H.; Wang, D.; and Bosch, N. 2025. Learning Behaviors Mediate the Effect of AI-powered Support for Metacognitive Calibration on Learning Outcomes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.
- Lempert, R. J. 2025. Decision Making under Deep Uncertainty and the Great Acceleration: The Role of Experts and Policy Analysis in a World in Transition. Expert Insights PE-A3789-1, RAND Corporation, Santa Monica, CA.
- Li, T.; Reiman, K.; Agarwal, Y.; Cranor, L. F.; and Hong, J. I. 2022. Understanding Challenges for Developers to Create Accurate Privacy Nutrition Labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–24. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.
- Lin, L.; Mu, H.; Zhai, Z.; Wang, M.; Wang, Y.; Wang, R.; Gao, J.; Zhang, Y.; Che, W.; Baldwin, T.; Han, X.; and Li, H. 2024. Against The Achilles' Heel: A Survey on Red Teaming for Generative Models. arXiv:2404.00629.
- Lu, J.; Holleis, T.; Zhang, Y.; Aumayer, B.; Nan, F.; Bai, H.; Ma, S.; Ma, S.; Li, M.; Yin, G.; Wang, Z.; and Pang, R. 2025. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 1160–1183. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.
- Madaio, M.; Kapania, S.; Qadri, R.; Wang, D.; Zaldivar, A.; Denton, R.; and Wilcox, L. 2024. Learning about Responsible AI On-The-Job: Learning Pathways, Orientations, and Aspirations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1544–1558. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Marchau, V. A. W. J.; Walker, W. E.; Bloemen, P. J. T. M.; and Popper, S. W., eds. 2019. *Decision Making under Deep Uncertainty: From Theory to Practice*. Springer Cham. ISBN 978-3-030-05252-2. Open access.
- Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; Ligett, K.; Lyons, T.; Manyika, J.; Niebles, J. C.; Shoham, Y.; Wald, R.; Walsh, T.; Hamrah, A.; Santarlasci, L.; Lotufo, J. B.; Rome, A.; Shi, A.; and Oak, S. 2025. Artificial Intelligence Index Report 2025. arXiv:2504.07139.
- McCaslin, T.; Alaga, J.; Nedungadi, S.; Donoughe, S.; Reed, T.; Bommasani, R.; Painter, C.; and Righetti, L. 2025. STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports. ArXiv:2508.09853 [cs].
- McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW): 72:1–72:23.
- McKinsey & Company. 2025. The State of AI: Global Survey 2025. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>.
- Mehandru, N.; Robertson, S.; and Salehi, N. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2016–2025. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.
- Menlo Ventures. 2025. 2025: The State of Generative AI in the Enterprise. <https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>.
- METR. 2025. What should companies share about risks from frontier AI models?
- Meyer, B. D. 1995. Natural and Quasi-Experiments in Economics. *Journal of Business & Economic Statistics*, 13(2): 151–161.
- Milli, S.; Belli, L.; and Hardt, M. 2022. Causal Inference Struggles with Agency on Online Platforms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 357–365. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.
- Mouton, C. A.; Lucas, C.; and Guest, E. 2024. The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. Technical report, RAND Corporation.
- Narayanan, A.; and Kapoor, S. 2025. GPT-4 and professional benchmarks: the wrong answer to the wrong question. New York State Assembly. 2025. RAISE Act, Assembly Bill A09449. https://assembly.state.ny.us/leg/?default_fld=&leg_video=&bn=A09449&term=2025&Summary=Y&Actions=Y&Memo=Y&Text=Y. Accessed: 2026-03-10.
- OpenAI. 2025. Preparedness Framework (Version 2). <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>. Accessed: 2026-05-20.
- Park, J. S.; Zou, C. Q.; Kamphorst, J.; Egan, N.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Liang, P.; Willer, R.; and Bernstein, M. S. 2026. LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. arXiv:2411.10109.
- Parker, C.; Scott, S.; and Geddes, A. 2019. Snowball Sampling. *SAGE Research Methods Foundations*. Publisher: SAGE.
- Paskov, P.; Byun, M. J.; Wei, K.; and Webster, T. 2025. Preliminary suggestions for rigorous GPAI model evaluations.
- Patwardhan, T.; Liu, K.; Markov, T.; Chowdhury, N.; Leet, D.; Cone, N.; Maltbie, C.; Huizinga, J.; Wainwright, C.; Jackson, F.; Adler, S.; Casagrande, R.; and Mardy, A. 2024. Building an early warning system for LLM-aided biological threat creation.

- Price, T. W.; Williams, J. J.; Solyst, J.; and Marwan, S. 2020. Engaging Students with Instructor Solutions in Online Programming Homework. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.
- Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. arXiv:2111.15366.
- Ratkovic, M.; Gschwend, T.; Wendering, L.; Bauer, K.; Kurella, A.-S.; Rittmann, O.; Sauter, M.; and Schwitter, N. 2025. Harnessing GPT for Enhanced Academic Writing: Evidence from a Field Experiment with Early-Career Researchers in the Social Sciences.
- Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Comanescu, R.; Akbulut, C.; Stepleton, T.; Mateos-Garcia, J.; Bergman, A. S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; Isaac, W.; and Weidinger, L. 2024. Gaps in the Safety Evaluation of Generative AI. In Das, S.; Green, B. P.; Varshney, K.; Ganapini, M. B.; and Renda, A., eds., *AIES (I)*, 1200–1217. AAAI Press.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022.
- Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. arXiv:2411.12990.
- Rostamzadeh, N.; Mincu, D.; Roy, S.; Smart, A.; Wilcox, L.; Pushkarna, M.; Schrouff, J.; Amironesei, R.; Moorosi, N.; and Heller, K. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1943–1961. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.
- Ruan, Y.; Dong, H.; Wang, A.; Pitis, S.; Zhou, Y.; Ba, J.; Dubois, Y.; Maddison, C. J.; and Hashimoto, T. 2024. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. arXiv:2309.15817.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.
- Salaudeen, O.; Reuel, A.; Ahmed, A.; Bedi, S.; Robertson, Z.; Sundar, S.; Domingue, B.; Wang, A.; and Koyejo, S. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. ArXiv:2505.10573 [cs].
- Scaramuzza, F.; Ferreira, R. C.; Quattrocchi, G.; Tamburri, D. A.; and van den Heuvel, W.-J. 2026. "Show Me You Comply... Without Showing Me Anything": Zero-Knowledge Software Auditing for AI-Enabled Systems. arXiv:2510.26576.
- Scaramuzza, F.; Quattrocchi, G.; and Tamburri, D. A. 2025. Engineering Trustworthy Machine-Learning Operations with Zero-Knowledge Proofs. arXiv:2505.20136.
- Schaeffer, R.; Schoelkopf, H.; Miranda, B.; Mukobi, G.; Madan, V.; Ibrahim, A.; Bradley, H.; Biderman, S.; and Koyejo, S. 2025. Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive?
- Schwartz, R.; Chowdhury, R.; Kundu, A.; Frase, H.; Fadaee, M.; David, T.; Waters, G.; Taik, A.; Briggs, M.; Hall, P.; Jain, S.; Yee, K.; Thomas, S.; Bhandari, S.; Duncan, P.; Thompson, A.; Carlyle, M.; Lu, Q.; Holmes, M.; and Skeadas, T. 2025. Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects. arXiv:2505.18893.
- Sekwenz, M.-T.; Wagner, B.; and Parkin, S. 2025. "It is unfair, and it would be unwise to expect the user to know the law!" – Evaluating reporting mechanisms under the Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 532–546. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1482-5.
- Shah, R.; Irpan, A.; Turner, A. M.; Wang, A.; Conmy, A.; Lindner, D.; Brown-Cohen, J.; Ho, L.; Nanda, N.; Popa, R. A.; Jain, R.; Greig, R.; Albanie, S.; Emmons, S.; Farquhar, S.; Krier, S.; Rajamanoharan, S.; Bridgers, S.; Ijtioye, T.; Everitt, T.; Krakovna, V.; Varma, V.; Mikulik, V.; Kenton, Z.; Orr, D.; Legg, S.; Goodman, N.; Dafoe, A.; Flynn, F.; and Dragan, A. 2025. An Approach to Technical AGI Safety and Security. ArXiv:2504.01849 [cs].
- Shanghai AI Lab; Chen, X.; Chen, Y.; Chen, Z.; Chen, Z.; Cui, H.; Duan, Y.; Guo, J.; Guo, Q.; Hu, X.; Huang, H.; Huang, L.; Li, C.; Li, J.; Lin, Q.; Liu, D.; Liu, X.; Liu, Z.; Lu, C.; Lu, X.; Qu, J.; Ren, Q.; Shao, J.; Shi, J.; Sun, J.; Wang, P.; Wang, W.; Xu, J.; Yan, L.; Yu, X.; Yu, Y.; Zhang, B.; Zhang, J.; Zhang, W.; Zheng, Z.; Zhou, T.; and Zhou, B. 2025. Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report. ArXiv:2507.16534 [cs] version: 2.
- Shevlane, T.; and Dafoe, A. 2020. The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 173–179. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.
- Strauss, I.; Moure, I.; O'Reilly, T.; and Rosenblat, S. 2025. The State of AI Governance Research: AI Safety and Reliability in Real World Commercial Deployment.
- Tong, X.; Mauriello, M. L.; Mora-Mendoza, M. A.; Prabhu, N.; Kim, J. P.; and Paredes Castro, P. E. 2023. Just Do Something: Comparing Self-proposed and Machine-recommended Stress Interventions among Online Workers with Home Sweet Office. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–20. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.
- UK AI Security Institute. 2024a. AI Safety Institute approach to evaluations.
- UK AI Security Institute. 2024b. Early lessons from evaluating frontier AI systems | AISI Work.

- U.S. AI Safety Institute. 2024. NIST AI 800-1 Managing Misuse Risk for Dual-Use Foundation Models.
- Vrijenhoek, S.; Daniil, S.; Sandel, J.; and Hollink, L. 2024. Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 573–584. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; Garcia-Gathright, J.; Olteanu, A.; Pangakis, N.; Reed, S.; Sheng, E.; Vann, D.; Vaughan, J. W.; Vogel, M.; Washington, H.; and Jacobs, A. Z. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. ArXiv:2502.00561 [cs].
- Wang, A.; Datta, T.; and Dickerson, J. P. 2025. Strategies for Increasing Corporate Responsible AI Prioritization. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 1514–1526. AAAI Press.
- Wei, K.; Ezell, C.; Gabrieli, N.; and Deshpande, C. 2024. How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1539–1555. Number: 1.
- Wei, K.; Guth, S.; Wu, G.; and Paskov, P. 2025a. Methodological Challenges in Agentic Evaluations of AI Systems. In *ICML Workshop on Technical AI Governance (TAIG)*.
- Wei, K. L.; Paskov, P.; Dev, S.; Byun, M. J.; Reuel, A.; Roberts-Gaal, X.; Calcott, R.; Coxon, E.; and Deshpande, C. 2025b. Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations. arXiv:2506.13776.
- Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.
- Wilson, C. 2024. The US Can Win Without Compromising AI Safety. *Tech Policy Press*.
- xAI. 2025. Grok 4.1 Model Card. Technical report.
- Yao, K.; and Juarez, M. 2026. SPRINT: Robust Model Attribution of Generated Images via Secret Pixel Reconstruction. arXiv:2508.05691.
- You, S. 2026. Verifiable Audit Mechanisms in AI Compliance Automation: Scalability. SSRN working paper. Available at SSRN: <https://ssrn.com/abstract=6547458>.
- Zhang, C. B. C.; Knight, C. Q.; Kruus, N.; Hausenloy, J.; Medeiros, P.; Li, N.; Kim, A.; Orlovskiy, Y.; Breen, C.; Cai, B.; Götting, J.; Liu, A. B.; Nedungadi, S.; Rodriguez, P.; He, Y. Y.; Shaaban, M.; Wang, Z.; Donoughe, S.; and Michael, J. 2026. LLM Novice Uplift on Dual-Use, In Silico Biology Tasks.
- Zhang, H.; Chen, P.; Yang, J.; Wu, Y.; Jiang, Z.; Xie, X.; You, W.; and Sun, L. 2025. IEDS: Exploring an Intelligent Embodied Design Space Combining Designer, AR, and GAI to Support Industrial Conceptual Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–25. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.
- Zhu, X.; Ye, Y.; Qiu, T.; Zhu, H.; Tan, S.; Mannan, A.; Michala, J.; Popa, R. A.; and Neiswanger, W. 2026. Auditing Black-Box LLM APIs with a Rank-Based Uniformity Test. arXiv:2506.06975.