

Measuring and Mitigating (Over)reliance to Build Human-Compatible AI

Lujain Ibrahim^{1*}, Katherine M. Collins², Sunnie S. Y. Kim³, Anka Reuel^{4, 5},
Max Lamparth⁴, Kevin Feng⁶, Lama Ahmad⁷, Prajna Soni², Alia El Kattan⁸, Merlin Stein^{1, 9},
Siddharth Swaroop⁵, Vishakh Padmakumar⁴, Ilia Sucholutsky⁸, Andrew Strait⁹, Diyi Yang⁴,
Q. Vera Liao¹⁰, Umang Bhatt¹¹

¹University of Oxford

²Massachusetts Institute of Technology

³Microsoft

⁴Stanford University

⁵Harvard University

⁶University of Washington

⁷OpenAI

⁸New York University

⁹UK AI Security Institute

¹⁰University of Michigan

¹¹University of Cambridge

Abstract

Large language models (LLMs) differ from previous technologies in functioning as collaborative “thought partners” that can engage fluidly in natural language across a range of tasks and increasingly act as autonomous agents on users’ behalf. As they come to shape consequential decisions from healthcare to personal advice, understanding how users rely on them, and when that reliance becomes miscalibrated, grows more urgent. This paper argues that measuring and mitigating *overreliance* must become central to LLM research and deployment. First, we consolidate risks from overreliance at both the individual and societal levels, including high-stakes errors, governance challenges, and long-term impacts on skill development and autonomy. Then, we explore LLM characteristics, system design features, and user cognitive biases that together raise serious and unique concerns about overreliance on LLMs in practice. We also examine historical approaches for measuring reliance, identifying three gaps and proposing three directions to improve measurement. Finally, we discuss mitigation strategies that can be pursued to ensure LLMs augment rather than diminish human capabilities.

Introduction

Large language models (LLMs) are increasingly trusted with consequential decisions in domains such as healthcare, education, and personal life (Zao-Sanders 2025a; Chatterji et al. 2025), where they support core cognitive functions like knowledge generation and reasoning (Collins et al. 2024b). However, our understanding of how humans interact with LLMs as thought partners—and increasingly as agents that act on their behalf—lags behind their rapid adoption. This

paper focuses on *overreliance*—when users rely on LLMs beyond their warranted capabilities, adopting incorrect outputs or inappropriately delegating decisions (Kim et al. 2023). We argue that measuring and mitigating overreliance must become a central focus in LLM research and deployment.

While the study of human reliance on technology is well-established, the unique characteristics of LLMs—their conversational nature, complex text-based outputs, and versatility—complicate traditional approaches to measuring and mitigating (over)reliance and its impacts. Their conversational nature means reliance unfolds across turns rather than at a single point of acceptance or rejection, making the source of influence hard to isolate. LLM outputs can also be only partially correct, or not verifiable at all, and users may adopt some elements while rejecting others. More broadly, the general-purpose nature of LLMs can foster an illusion of universal competence (Messeri and Crockett 2024), masking domain- or context-specific limitations and inconsistencies. This versatility also enables reliance to extend to novel domains of use, such as personal advice and moral reasoning (Tennant et al. 2026; Cheng et al. 2025; Ibrahim et al. 2026), and to new modes of interaction, such as the delegation of various actions (e.g., making purchases or sending emails) to LLM agents.

If overreliance is not measured, and steps are not taken towards its mitigation, LLMs, however capable, may fail to be human-compatible, preventing them from reaching their full potential in serving real users. They will be neither maximally helpful nor maximally harmless as many AI practitioners strive to build (Huang et al. 2025; Sharma et al. 2026). By systematically measuring and mitigating overreliance, we can better ensure LLMs augment human capabilities and enhance user autonomy, rather than diminish,

*Corresponding author: lujain.ibrahim@oii.ox.ac.uk

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

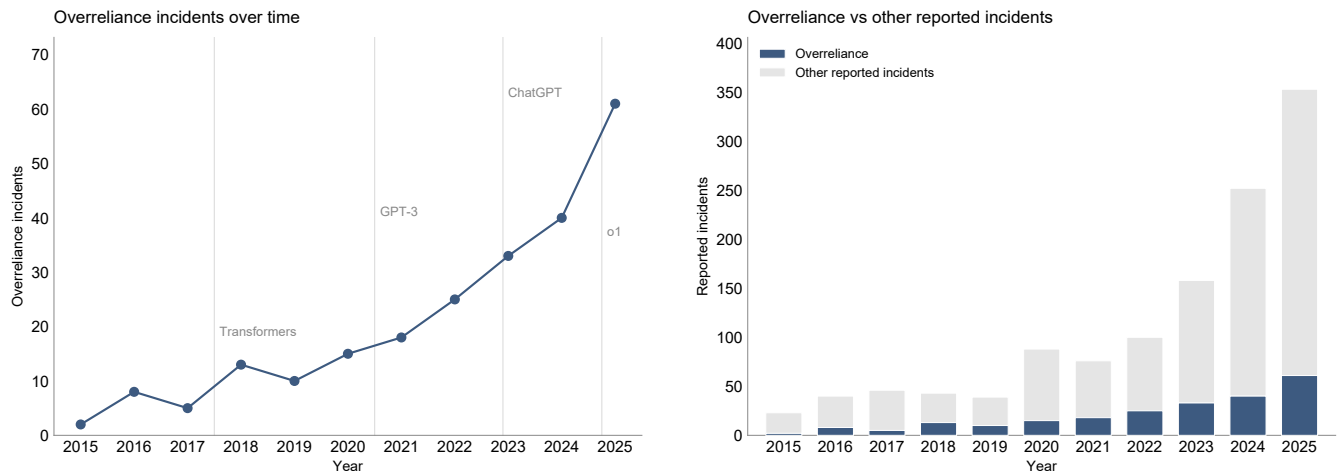


Figure 1: Reported incidents involving overreliance on AI systems, 2015–2025. Incidents were extracted from the AI Incident Database ($N = 1,339$) and labeled for whether they involved overreliance using unanimous agreement across three LLM judges (GPT-4o-mini, GPT-4.1, Claude 4.5 Sonnet). Overreliance was defined as cases where humans relied on AI systems (not exclusive to LLMs) without adequate verification, supervision, or critical evaluation. This figure is intended as illustrative rather than as evidence of a trend, as the AI Incident Database is a voluntarily curated collection and thus its contents reflect reporting practices and biases. Vertical lines indicate major LLM capability milestones. On the left: absolute number of incidents classified as overreliance per year. On the right: overreliance incidents (dark blue) stacked against other reported incidents (light gray) per year. LLM judge ratings were validated by one of the study authors (accuracy 0.78, Cohen’s $\kappa = 0.53$; see Appendix).

displace, or homogenize them. To that end, this paper contributes the following:

- Grounding the study of overreliance on LLMs with an overview of the **history of reliance research** across disciplines
- Articulating key **risks and potential impacts** of overreliance across individual and societal levels and near- and long-term horizons
- Identifying gaps in existing **measurement** approaches that arise from the unique characteristics of LLMs, and proposing new directions for measuring overreliance in human-LLM interactions
- Proposing promising directions for **mitigation** strategies for multiple research communities to address (technical AI research, human-computer interaction, and policy).

A Brief History of Reliance Research

Human reliance on technological systems has been a subject of study in many fields, including human–computer interaction, cognitive science, psychology, computer science, and management science. Parasuraman and Riley (1997)’s typology of automation use categorizes misuse (overreliance) and disuse (underreliance) as opposite failures of human-machine interaction, as both forms can degrade interaction quality and outcomes. Overreliance, the focus of this paper, is commonly defined as either adopting a system output when that output is wrong, or delegating to a system when such delegation is undesirable (Kim et al. 2023). Specifically, overreliance is defined as a behavior, not a “feeling nor an attitude, but the actual action con-

ducted” (Schemmer et al. 2023). For example, an act of overreliance is committed when a programmer implements or accepts LLM-generated code suggestions that contain security vulnerabilities, as they defer to the system even if they have the information or expertise to act differently. Underreliance, at the other end of the spectrum, could arise when users overestimate their own capabilities or distrust machines due to past errors or pre-existing discomfort with automation (Dietvorst, Simmons, and Massey 2018; Parasuraman and Manzey 2010). Dubbed “algorithm aversion” (Dietvorst, Simmons, and Massey 2015), individuals may ignore accurate recommendations or avoid using the system altogether, forfeiting potential benefits.

Overreliance is frequently distinguished from trust, a related but distinct construct (Dzindolet et al. 2003; Hou and Jung 2021; Zerilli, Bhatt, and Weller 2022). While trust represents a psychological state of willingness to be vulnerable based on positive expectations of others’ actions (O’Neill 2018), overreliance often manifests as a behavioral consequence of excessive trust (Vereschak, Bailly, and Caramiaux 2021). The relationship between trust and reliance is not linear (Kahr et al. 2024; Lee and See 2004). While trust guides the intention to rely, the action of reliance can also be influenced by other expectations and constraints such as saving time, reducing workload, avoiding risks, and exploring capabilities.

Overreliance behaviors are often described as errors of commission, and they typically emerge without intent to offload responsibility, marking overreliance as an unintended outcome (Passi and Vorvoreanu 2022). To explain it, some researchers emphasize cognitive mechanisms behind

overreliance, such as automation bias, complacency bias, and cognitive offloading (Rastogi et al. 2022). Others focus on design or systemic factors. For instance, systems like current LLMs that are unable to reliably produce uncertainty estimates, frictionless interfaces, or organizations that treat AI systems as authoritative, can all foster overreliance by signaling that oversight is not possible or is unnecessary (Cummings 2006). Despite differences in emphasis across disciplines, definitions converge on a few key points that we also center in this work: (1) overreliance is typically unintentional, (2) overreliance emerges from misaligned trust or expectations between the user and the system, and (3) overreliance produces measurable failures in judgment or oversight (Schemmer et al. 2023).

Risks from (Over)reliance on LLMs

As shown in Figure 1, overreliance consistently appears among reported AI incidents over the past decade (though it is important to note that incident databases reflect reporting and curation practices rather than just underlying incidence). Here, we expand on these impacts of overreliance across domains and timescales.

Individual risks

Near-term: high-stakes errors In some instances, users may rely on AI systems for real-time information, or as inputs for decision-making, which can result in immediate errors or harm. This could occur in both personal and professional use. For example, an incorrect suggestion from an LLM used for a health-related decision could lead to a misdiagnosis or unsafe treatment plan, both in cases of self-diagnosis and misdiagnosis by a professional (AIID 2024; Choudhury and Chaudhry 2024). Similarly, unreliable information can be generated by LLMs in legal contexts, such as in a recent incident where lawyers submitted a court brief that unknowingly included six non-existent LLM-generated case citations (Castel 2023). Furthermore, developers relying on coding agents that can write and deploy code without proper verification may end up deploying applications with critical security vulnerabilities. Such cases highlight that even skilled professionals may overrely on incorrect outputs that sound plausible, especially when there is not adequate time or resourcing available to critically evaluate them (Nishal, Sinchai, and Diakopoulos 2024), or when there is limited education about where systems may be overconfident or error-prone (Long and Magerko 2020; Vasconcelos et al. 2023; Ibrahim, Ghassemi, and Alhanai 2023). As AI systems become more autonomous, self-improving, or deployed in open-ended environments, their brittleness and goal-alignment issues may lead to unprecedented failures.

Long-term: dependence and deskilling As users regularly defer to AI systems for cognitive tasks, they risk losing both skills and motivation for independent problem-solving (Gerlich 2025), potentially creating a negative feedback loop of diminishing ability to calibrate appropriate trust (Chan et al. 2023). One can argue that narrow deskilling, such as losing mental arithmetic speed due to

calculators, is less concerning because the tool is highly reliable, the skill is well-scoped, and the stakes are typically low. However, LLM use risks inducing *broad* deskilling across foundational and interconnected cognitive capacities (e.g., knowledge acquisition, critical reasoning, and emotional processing). Cases of both narrow and broad deskilling are increasingly documented across domains including general problem-solving capacity (Zhai, Wibowo, and Li 2024), civic engagement (Kudina and de Boer 2024), and healthcare practice (Choudhury and Chaudhry 2024; Budzyń et al. 2025). Another critical domain is emotional dependence on AI systems. Unlike task-based overreliance, this involves users developing inappropriate attachments to LLMs for emotional support, companionship, or relationship simulation, potentially displacing human connections (Ibrahim et al. 2026; Earp et al. 2025). There may also be a compounding risk from deskilling and overreliance: dubbed the “ironies of automation” (Bainbridge 1983), the more advanced the technical capability, the more skilled users may need to be to detect and correct errors, but (over)reliance on these capabilities makes the required skill more challenging to develop and preserve.

Societal risks

Near-term: governance challenges At an institutional level, reliance on LLMs can expose organizations to operational failures, decision-making biases, and cascading risks, particularly when human oversight is reduced or overlooked. For example, developers accepting flawed AI-generated code could lead to service disruptions, especially if there are no mechanisms in place to identify developers who fail to critically evaluate outputs (Pu et al. 2025; Spatharioti et al. 2025). Governance becomes especially challenging in sectors such as critical infrastructure, healthcare, and public safety, where multiple organizations may procure and deploy the same AI systems (Balayn et al. 2024). Without sufficient oversight, reliance patterns across institutions can amplify failures and cause widespread disruptions (Cui, Li, and Zhang 2022). Another example is in financial markets, where reliance on AI-generated news or stock analysis without proper regulatory oversight could lead to flash crashes if multiple systems react similarly to the same misinterpreted input (Azzutti 2024). The governance challenge here lies in ensuring that AI systems comply with regulatory frameworks in these sectors, especially in areas where biases or errors could have significant legal and reputational consequences (Bhatt and Sargeant 2024).

Long-term: influencing norms and homogenization Over time, AI use will likely generate new norms that shape behaviors, beliefs, and creative practices, essentially acting as new sources of “culture” or agents of social learning (Brinkmann et al. 2023; Collins, Bhatt, and Sucholutsky 2025). One concern is that as organizations adopt LLMs for content creation (e.g., news articles, reports, policy drafts) or decision support, AI-generated outputs could unintentionally shape human discourse and norms. For instance, researchers have noted that reliance on AI-generated content risks creating an “echo chamber” that stifles novel

Domain	Short-term	Long-term
Healthcare	Individual: Medical doctor follows incorrect AI diagnosis Institutional: Hospital systems implement AI diagnostics without adequate verification protocols	Individual: Healthcare professionals experience cognitive deskilling and atrophy of diagnostic abilities Institutional: Medical departments restructure workflows around AI systems, reducing human oversight
Personal advice	Individual: User accepts sycophantic or biased relationship/career advice without critical evaluation Societal: Communities adopt similar AI-generated advice, creating homogenized decision patterns	Individual: Self-worth becomes derived from AI companion approval Societal: Social norms shift toward AI validation of personal choices

Table 1: Examples of short- and long-term impacts of AI reliance across two domains: healthcare and personal advice.

ideas and undermines the diversity of thought (Csernatoní 2024). This effect can scale when individual reliance influences institutional-level outputs and decision-making. For example, randomized control trials show that LLM-assisted writers generate less diverse ideas and converge on similar phrasings, even after the AI is removed (Ashkinaze et al. 2024; Abdulhai et al. 2026). That can flatten institutional and societal cultural practices, a phenomenon researchers have termed “algorithmic monocultures” or “culture flattening” (Chayka 2025; Kleinberg and Raghavan 2021), which may become especially problematic as systems directly engage with us in new forms of “collaborative cognition” (Oktar et al. 2025).

Factors Influencing (Over)reliance on LLMs

To address these diverse risks, it is necessary to understand the factors that shape reliance and may lead to overreliance. Previous work has proposed utility frameworks for modeling reliance decisions (Guo et al. 2024; Wang, Lu, and Yin 2022), suggesting that reliance is influenced by factors affecting (1) perceived model correctness or capability given the task (if relying), (2) perceived self-correctness or capability (if not relying), and (3) evaluation of outcome payoffs (e.g., stakes and risk tolerance). We organize this section around model, system, and user factors to facilitate our later discussion of targeted mitigation strategies. *Model characteristics* primarily influence perceived model correctness; *system design* can affect perceived model correctness and payoff evaluations; and *user factors* can shape perceived self-correctness and payoff evaluations.

Model characteristics

Several characteristics of LLMs can contribute to overreliance, most fundamentally through the anthropomorphic qualities of their natural language output. Language models simulate human-like conversational patterns, e.g., through asking follow-up questions or expressing helpfulness and empathy, influencing users’ perceptions of their competence, trustworthiness, and alignment (Abercrombie et al. 2023; Ibrahim et al. 2025; McKee, Bai, and Fiske 2024). In some cases, users may develop an emotional attachment to models which could lead to increased reliance across both

emotional and task-based domains, though this relationship remains an important area for future research (Manzini et al. 2024). Language models are also capable of expert mimicry, e.g., through citations and structured responses, assertive phrasing, or confident tone, which can further signal certainty and reinforce perceived authority (Metzger et al. 2024). Notably, language models can communicate hallucinations or fabricated content as factual statements, using the same fluent, human-like, and often confident language, making errors difficult to detect. Alongside social and expert mimicry, the tendency of models to be overly agreeable or sycophantic can reinforce users’ existing beliefs and preferences, creating a feedback loop that discourages critical evaluation (Cheng et al. 2025; Sharma et al. 2023; Ibrahim, Hafner, and Rocher 2025). Collectively, these characteristics can contribute to users conflating stylistic polish with epistemic reliability, as current models often default to confident, fluent outputs and are simultaneously unable to reliably express uncertainty that could prompt critical evaluation (Zhou et al. 2024; Yin et al. 2023).

System design

Research in AI and human-computer interaction has investigated a wide range of interactive affordances that mediate human trust and reliance on AI (Amershi et al. 2019). For example, the availability, or lack thereof, of (un)certainly indicators that communicate to users the likelihood of correctness of AI outputs has been shown to impact overreliance patterns (Kim et al. 2024; Ma et al. 2023; Zhou et al. 2024). Notably, while emerging research has shown promise in token-level methods, sample-based methods (Shorinwa et al. 2024), and multicalibration (Detommaso et al. 2024), novel challenges still remain in reliably computing (un)certainly of LLMs compared to their predictive predecessors. Another frequently studied system affordance is the presence of model explanations, such as feature-importance explanations that indicate how the model weighs the features of the input to arrive at its predictions, or in the case of LLMs, “chains of thought” or reasoning traces. Contrary to the common expectation that transparency should improve human oversight, many empirical studies find that explanations in fact can increase overreliance

 Model	↔	 System	↔	 User
Anthropomorphic language Sycophancy Certainty expressions		Explanations Interface design Frictions		Cognitive capacity User experience Cognitive biases

Table 2: Selected factors influencing (over)reliance on LLMs.

(and simultaneously reduce underreliance) (Zhang, Liao, and Bellamy 2020; Chen et al. 2023; Bansal et al. 2021; Kim et al. 2025). The reason is that the mere presence of explanation can be associated with model competence, and potentially disrupt people’s own reasoning process. On the other hand, *adaptive explanations* which can tailor explanations by task context or levels of confidence, were found to be more effective in calibrating reliance (Lai et al. 2023b; Robrecht, Rothgänger, and Kopp 2023; Turchi et al. 2024; Ma et al. 2023). Interface designs, such as added friction or selective task delegation in multi-step agentic trajectories, have also been shown to influence reliance behaviors (Feng et al. 2024; He, Demartini, and Gadiraju 2025).

User cognition and experience

Several user factors may explain why a user overrelies on an LLM. Users with a reduced “need for cognition” (Cacioppo and Petty 1982) are more likely to overrely (Bućinca, Malaya, and Gajos 2021), while a user’s disposition towards trusting and their general intuitions about the task at hand (e.g., their level of expertise) can influence baseline reliance behavior (Chen et al. 2023; Hoff and Bashir 2015). However, even without a predisposition to trust or with low confidence, users may still be swayed to overrely due to inherent cognitive limitations. Humans are constrained by time and cognitive capacity (Griffiths 2020), which is often addressed through “resource-rational” analyses (Lieder and Griffiths 2020). These limitations suggest that when a task demands more than can be processed under our limited “compute budget,” users may default to relying on a model’s output without proper verification (Lee et al. 2025; Sarkar 2024; Simkute et al. 2025). Propensity towards reliance is further shaped by user experience. Emerging research highlights a potential relationship between chatbot usage and reliance on model outputs, for example in social or affective contexts (Fang et al. 2025; Phang et al. 2025). While humans have the ability to update beliefs in light of new evidence (Griffiths, Chater, and Tenenbaum 2024; Tenenbaum et al. 2011), this updating process can go awry, with individuals giving undue weight to particularly salient experiences (Chater et al. 2020; Tversky and Kahneman 1974). In conversations, humans often make inferences about what is not said based on the structure of the conversation and come to these conversations with strong prior expectations about how they will unfold (Grice 1975). These expectations shape reliance decisions, with recent research showing that humans may struggle to generalize LLM successes and failures across domains, often miscalibrating their mental models of model behavior (Kelly et al.

2023; Steyvers and Kumar 2024; Vafa, Rambachan, and Mullainathan 2024; Yu et al. 2026).

Measuring (Over)reliance on LLMs

Long before “reliance on AI” emerged as a research area, reliance has been a subject of extensive measurement research across fields (Lee and Moray 1992, 1994; Lee and See 2004; Manzey, Reichenbach, and Onnasch 2012). Much of this work investigates reliance on human advice using the *judge-advisor framework* (Bonaccio and Dalal 2006), where a “judge” receives advice from an “advisor” and must decide what to do with the advice. Here, reliance is typically measured as “advice acceptance,” where studies have investigated various factors that influence this acceptance (Harvey and Fischer 1997; Snizek and Van Swol 2001; Van Dongen and Van Maanen 2013).

Recently, reliance on AI has been studied in AI-assisted decision-making scenarios following the same judge-advisor framework, where AI serves as the “advisor” and provides advice for the human “judge.” These studies, which surged around 2020, typically investigate discriminative models that predict a discrete label (Bućinca, Malaya, and Gajos 2021; Lai et al. 2023a; Zhang, Liao, and Bellamy 2020; Yin, Vaughan, and Wallach 2019). While a few recent studies investigate reliance on LLMs specifically, they focus on limited scenarios of making binary true/false judgments of factual statements (Kim et al. 2024, 2025; Si et al. 2023). This kind of setup allows the treatment of reliance as a binary decision of accepting or rejecting the whole or a part of AI output in the final answer of the human “judge” (Hullman, Kale, and Hartline 2025; Lai et al. 2023a).

However, this conceptualization is an approximation that does not distinguish between situations where people’s own judgment agrees with the AI output from ones where there is explicit reliance. Thus, other work instead measures reliance as “switch fraction,” which quantifies how often people change their initial answer to adopt the AI advice (He, Kuiper, and Gadiraju 2023; Lu and Yin 2021; Yin, Vaughan, and Wallach 2019; Zhang, Liao, and Bellamy 2020), or “weight of advice” (WoA), which measures the degree to which people shift their answer toward the AI advice on a continuous scale (Logg, Minson, and Moore 2019). Amongst these measures, only WoA can be extended to measure reliance in a non-discrete context, such as in a numerical estimation task where one changes their own score to be closer to AI’s score (Bo, Wan, and Anderson 2024).

While the above-mentioned measures reflect how people descriptively rely on AI outputs, the measurement of *appropriate* reliance is a more complex challenge. Appropriate

ate reliance simultaneously requires relying on AI when the AI advice is desirable or helpful (*not underrelying*, or correct AI reliance) and not relying on AI when the AI advice is undesirable or unhelpful (*not overrelying*, or correct self-reliance). Past research often examines the two components separately by measuring reliance on correct and incorrect AI outputs respectively (Kim et al. 2024; Wang and Yin 2021), or utilizes metrics that reflect the sum of the two (Wang and Yin 2021), the ratio of the two (Bo, Wan, and Anderson 2024), or treats them as a two-dimensional metric (Schemmer et al. 2023). Some work bypasses quantifying overreliance or underreliance per se, but instead takes an outcome-oriented view by measuring, at an aggregate level, whether the human-AI joint outcome (e.g., joint decision accuracy) is better than that of the human or AI alone (Yin, Vaughan, and Wallach 2019; Zhang, Liao, and Bellamy 2020).

Why existing measures break down for LLMs

As outlined above, traditional approaches to measuring reliance focus on experimental settings where humans interact with discriminative models that predict a discrete label. However, such approaches falter in the dynamic, conversational settings typical of human-LLM interactions. LLMs differ fundamentally from prior AI systems: their **complex text-based outputs**, **interactive use**, and **general-purpose nature** complicate how reliance, and thus overreliance, are measured.

Complex text-based outputs Traditional reliance metrics such as agreement and switch fraction reviewed above presume binary correctness and a known ground truth—assumptions rarely met in open-ended generative tasks (Passi, Dhanorkar, and Vorvoreanu 2024). Crucially, these metrics fail to capture the granular nature of reliance with LLM outputs, where the *unit of reliance* can vary non-discretely across token, sentence, and topical levels. Users may rely on specific semantic elements while rejecting others within the same output, or trust the topical direction while questioning particular factual claims. This multi-layered reliance structure means LLM outputs may often be partially desirable, and users may engage with different components of them at varying levels of trust and critical evaluation. Additionally, how users engage with these complex LLM outputs can vary by task types and user goals (Chevi et al. 2025).

Interactive use LLMs undermine the foundational assumption of *output-based* metrics in prior work. Rather than a single-shot interaction where a user accepts or rejects an output with fixed input, LLM usage involves iterative prompting and refinement, in which (over)reliance may unfold over time and be reflected in both how the user engages with LLM outputs and how they formulate the inputs. As a result, paradigms like the judge-advisor framework (Bonaccio and Dalal 2006; Kim et al. 2024; Lu and Yin 2021), which require clean separations between user and AI contributions, struggle to capture the nature of LLM use. Furthermore, unlike earlier systems with traceable recommendations, LLMs operate through continuous back-and-forth exchanges, making it challenging to pinpoint the exact source

of influence. The influence of LLMs also goes beyond direct suggestions, subtly shaping framing, language, and ideation over dialogue turns. These challenges are amplified when LLMs operate as autonomous agents, where users delegate actions and may not observe intermediate outputs at all, making reliance concerned more with process rather than any single output.

General-purpose nature LLMs are used across a wide spectrum of tasks, from information retrieval to creative writing, each with fundamentally different notions of what constitutes a good output or a good outcome. This diversity challenges traditional concepts of “appropriate reliance.” In domains like coding or information retrieval, outputs can be evaluated against established standards of correctness. However, in areas like emotional support, relationship advice, or career guidance, which are growing in popularity, there is no clear benchmark for “correct” outputs, and what counts as a successful outcome is also subjective (Zao-Sanders 2025b). In domains like education, while output correctness may be available, a good outcome is not typically defined in terms of productivity, as a student who obtains a correct output may still have failed to learn. Without stable criteria at either level, determining when reliance is appropriate becomes more challenging, even as users engage with LLMs in ways that affect their autonomy and well-being (Banker and Khetani 2019).

Toward better measurement of (over)reliance on LLMs

Given the shortcomings of existing measures for capturing reliance behaviors in human-LLM interactions, we outline three directions for developing better measures, each with concrete operationalizations. Table 3 summarizes how LLM characteristics create measurement challenges and maps these to specific proposed approaches.

Go beyond single-point measures LLMs produce more complex outputs than traditional machine learning models. For example, it is common for an LLM output to be highly fluent and factually accurate but have gaps in logic. Take a simple decision-making task as an example: a user has to decide between option A and option B. While traditional AI models would recommend a single option (A or B), LLMs may provide supporting arguments for both options, refrain from providing a recommendation, or recommend a single option but with varying arguments and degrees of confidence. In this example, how should reliance be measured? If a user agrees with an LLM’s argument for its recommendation but ultimately does not follow that recommendation, did the user rely on the LLM or not? What about vice versa? In other more open-ended and generation-based tasks (e.g., writing, coding, planning), single-point measures, such as agreement, may again fall short, as LLMs typically do not provide a single recommendation. For these tasks, measuring the extent to which a user adopts the LLM’s output (e.g., the number of words, tokens, or suggestions) may be a more suitable measure of reliance (Vasconcelos et al. 2023). For example, metrics like edit distance between a user’s draft and the final collaboration output, or computed semantic

LLM characteristic	Measurement challenge	Proposed metric(s)	Data source
Complex text-based outputs	Varying units of reliance (token, sentence, topic)	Edit distance, semantic similarity at multiple granularities	Pre/post output comparison, embeddings
Interactive use	No discrete decision points, reliance unfolds over time	Session-level adoption trajectory, turn-by-turn influence scores, counterfactual simulation	Interaction logs, conversation transcripts
General-purpose nature	Context-dependent success criteria, subjective tasks with no ground truth	Outcome-based assessment, goal completion assessments	Task completion data, delayed follow-up, surveys, longitudinal tracking

Table 3: Measurement framework: LLM characteristics, associated challenges, and proposed metrics.

similarity scores at different levels (e.g., sentence and paragraph levels) can capture reliance at varied granularities. For multi-turn interactions, tracking how reliance evolves across turns (including whether early LLM suggestions anchor later user decisions) may reveal patterns invisible to single-point measures and how task effort is distributed between users and AI (Padmakumar et al. 2026). These methods can also follow other calls for richer ways of evaluating LLMs and human-LLM interaction (Ying et al. 2025; Ibrahim et al. 2024).

Shift from outputs to outcomes In traditional reliance measurement, there is often a clear notion of “good” outputs, typically defined as correct classifications according to ground truth labels, based on which over-, under-, and appropriate reliance are defined. For LLM use in verifiable domains, like factual information retrieval, appropriate reliance can be similarly assessed against a ground truth. However, LLMs are increasingly used in subjective, non-verifiable domains, where there is no stable notion of “correct” output. Thus, this requires a shift in focus from the *outputs* of LLMs to the *outcomes* that result from user interactions with these outputs (e.g., did the user achieve their goal? Was user well-being or autonomy preserved?) when assessing the appropriateness of reliance. For a given task (e.g., producing a satisfactory piece of writing or code, or deciding on a course of action), “appropriate reliance” may be defined as relying on the LLM and achieving the desired outcome; “overreliance” as relying on the LLM and not achieving the desired outcome; and “underreliance” as not relying on the LLM and not achieving the desired outcome, when only relying on the LLM would have given the desired outcome. Importantly, there is subjectivity in defining what is considered a desired outcome, and process may diverge from outcome: reliance may be epistemically inappropriate yet result in a good outcome, or appropriate yet result in a bad one. For example, consider a user who requests career advice from an LLM, engaging appropriately and verifying its claims, yet the outcome of their career decision is poor for external reasons unrelated to how they relied. There must also be considerations of what is an appropriate time horizon for assessing outcomes, e.g., it could be immediately for factual tasks, but potentially days or weeks later for advice-seeking or planning tasks where consequences unfold over time.

Employ additional measures to holistically understand reliance

While adopting and adapting commonly used behavioral measures of reliance (e.g., agreement) and engaging with historical research on reliance will remain critical, employing additional measures is necessary to gain complementary in-depth insights to specific human-LLM interaction contexts. For example, when studying reliance on LLMs in information-seeking contexts, additional tracking of users’ source clicking behavior and follow-up questions can provide deeper insights into when and how much users rely on LLMs and what factors shape their reliance (Kim et al. 2024, 2025). Other common additional measures include users’ self-reported confidence in their final answer and evaluation of AI outputs: increased confidence may indicate increased reliance, and this may be miscalibrated (Vodrahalli, Gerstenberg, and Zou 2022). Additionally, increased overreliance may lead to less learning and enjoyment of the task, which can also be measured (Glikson and Woolley 2020). For affective use or personal advice applications of LLMs, surveys and behavioral measures of user well-being, such as self-reported mood surveys, stress or anxiety scales, and behavioral indicators of decision satisfaction, can also be employed (Cao and Huang 2022). Overall, analyses of user interaction logs, such as investigating how users frame requests, what they choose to delegate, and when, will be central to surfacing implicit patterns of reliance that behavioral metrics and self-report measures may miss.

Mitigating Overreliance on LLMs

Once we can better measure overreliance, the next challenge is mitigation, for which we outline several directions below.

Model-level mitigations

Attending to the linguistic and stylistic choices in language model training is important for managing overreliance at the model level. Current LLMs exhibit several patterns that may influence user perceptions of system capabilities and confidence. First, human-like language features, such as empathetic responses and social pleasantries, can lead users to attribute greater agency and competence to systems (Abercrombie et al. 2023; Ibrahim et al. 2025). Second, LLMs often express high certainty: plain statements without epistemic markers are perceived as confident assertions, which

may cause users to overestimate response reliability (Zhou et al. 2024). These patterns appear to emerge partly from training processes. Existing work suggests that RLHF can amplify sycophancy (Sharma et al. 2023), reward modeling tends to favor certainty over accuracy, and human raters often prefer confident-sounding responses regardless of correctness (Zhou et al. 2024). LLMs also appear to assume users are more rational than they typically are, leading to outputs that do not account for known cognitive biases in how users interpret and act on information (Liu et al. 2024). Addressing these patterns may require targeted interventions. Improving models to accurately convey their uncertainty, in natural language (e.g., “I’m not sure, maybe...”) or other forms (e.g., confidence scores), could help users calibrate reliance (Kim et al. 2024; Vasconcelos et al. 2025). Future research might explore modified reward functions that account for unwarranted certainty and human feedback protocols that encourage more balanced (un)certainty expressions. We also encourage that any modifications that are made to models are appropriately communicated if receiving feedback from users, e.g., with FeedbackLogs (Barker et al. 2023), to ensure that users are aware when they may want to explicitly adjust their expectations for models.

System-level mitigations

Interface design choices that incorporate cognitive forcing functions and visual cues can help mitigate overreliance. For example, priming statements that clarify the model’s goals and limitations can set expectations of model capabilities (Collins et al. 2024a; Pataranutaporn et al. 2023). Mixed-initiative controls, where both the user and AI suggest, produce, evaluate, and modify outputs in response to each other, can help establish feedback loops that promote balanced interaction (Deterding et al. 2017; Feng et al. 2024; Horvitz 1999; Subramonyam et al. 2024). AI-assisted autocomplete can allow users to phrase queries more effectively, bridging the gap between user intent and system capabilities (Subramonyam et al. 2024). Friction mechanisms, like an extra click or disclaimers, can also slow users down at critical decision points, further reducing reliance on AI outputs (Collins et al. 2024a; Hunter et al. 2024; Naiseh et al. 2021; Weisz et al. 2024). Finally, visual cues like typing indicators or word-by-word output, which contribute to the illusion of human-like behavior, can be revised or minimized depending on the use case (Abercrombie et al. 2023). These system-level interventions, rather than modifications to the model itself, are even more important for LLM agents that can act on users’ behalf in virtual or physical space. Agentic use involves not just individual outputs, but extensive plans and actions taken before a user sees any result. Early evidence shows mixed results for “obvious” interventions, where displaying an agent’s reasoning traces can increase overreliance (Grunde-McLaughlin et al. 2026). Designing and testing system-level interventions that meaningfully reduce overreliance on LLMs remains an open challenge.

User-level mitigations

Educating users about AI systems, how they work, their capabilities, and their limitations, can empower individuals to

make more informed decisions about when to rely on AI outputs (Chen et al. 2025; Long and Magerko 2020). Such interventions should ideally cater to users’ existing knowledge, be presented in an engaging manner, and be broadly accessible. For example, recent AI literacy efforts have utilized games as a medium to communicate to younger users (Csapo et al. 2019). And, recent studies have found users often overrely on LLMs and are unaware of how their outputs could be flawed (e.g., LLMs can provide unfaithful and inconsistent explanations) (Kim et al. 2024, 2025); correcting such misconceptions could be engineered as part of AI literacy efforts. Further, while substantial effort has gone into benchmarking AI systems’ performance on various tasks, these benchmarks may not be immediately interpretable to the general public (e.g., what does an MMLU accuracy of 85% mean to an everyday user?) and warrant simplified summaries around expected performance. For instance, AI literacy efforts could contribute reports detailing expected model behavior to help correct for errors in our “generalization functions” (Vafa, Rambachan, and Mullainathan 2024).

Discussion

We close by addressing several tensions that arise in discussions of reliance on LLMs, each of which informs why we view measurement as a central and urgent research priority.

The current lack of empirical evidence

A reasonable starting concern is that (over)reliance on LLMs may not be a pressing issue in practice, given the general lack of empirical evidence. In other words, existing instances of harm from inaccurate LLM outputs or dependence on LLM-based systems are often isolated cases. However, we argue that limited empirical evidence at this time is not indicative of *no* issue of (over)reliance.

Rather, the current lack of evidence can be attributed, in part, to the absence of appropriate methods for measuring reliance in user-LLM interactions—as the saying goes, “what is not measured cannot be improved” (Kim et al. 2024, 2025). It is also due to barriers to accessing usage logs, whether of AI chatbots or agents, that would reveal how people use these systems and how they prompt them for various requests. Currently, this data is largely only available to AI providers. Controlled experiments offer an alternative, but are costly, especially if longitudinal, and may require expertise in the experimental sciences which does not always overlap with LLM expertise.

Importantly, when deployment occurs outside of controlled experimental settings, we cannot afford to wait for problematic or prevalent cases to emerge as evidence before taking action, especially for long-term harms such as deskilling, the erosion of critical thinking abilities, or emotional dependence (Casper, Krueger, and Hadfield-Menell 2025), which may be difficult to detect until they are deeply entrenched. These concerns are heightened as agentic capabilities and use increase and more users delegate actions to LLM agents, where the consequences of (over)reliance may be immediate and difficult to reverse. LLMs in their popular “chatbot” and “agent” forms remain early in their technological lifespan (circa 2022), and the identified instances of

harms, even if currently limited, should be viewed as important signals to intervene and catalyze work on measuring and mitigating potential risks (boyd 2025). Efforts to collect usage data in the wild, to require provider release of usage data in some form, to build post-deployment monitoring pipelines, and to develop privacy-preserving methods supporting such efforts will all be needed. Broader interdisciplinary collaboration will be central to collecting evidence on AI reliance.

The role of advancing AI capabilities

A second tension concerns whether overreliance is a transient problem. One might expect that as AI capabilities and alignment become increasingly sophisticated, overreliance concerns will naturally diminish: with hypothetically “superhuman” AI that rarely makes mistakes, traditional cases of overreliance would become increasingly rare. Paradoxically, however, as AI becomes more capable, detecting when these systems err becomes more difficult, as their outputs appear increasingly convincing. Even if overreliance on incorrect outputs becomes less frequent, the broader reliance harms we identify, such as deskilling, dependence, and disempowerment, may intensify rather than diminish. A highly capable system gives users more reason to defer and less incentive to maintain independent judgment, even when doing so remains important for their long-term well-being and agency. Overall, as capabilities accelerate, these systems will still require human oversight for legal and societal reasons (Bhatt et al. 2025; Steyvers et al. 2022). Advancing AI capabilities thus makes measuring reliance *more* essential, not *less*, to deploying AI for the benefit of all.

Societal vs. technical interventions

A final tension concerns where interventions should be located. One view holds that addressing overreliance through technical improvements or human-AI interaction design will ultimately prove fruitless, because the core challenges exist at societal and institutional levels. We agree that societal resilience and regulation are a critical piece of the puzzle, but contend that they hit a ceiling without complementary technical advances in measurement and system design. Regulations, for instance, cannot be effectively implemented without precise metrics to determine when and where overreliance is occurring. Educational and user empowerment initiatives may similarly face diminishing returns when systems themselves are designed in ways that encourage reliance (Lu et al. 2024). A comprehensive approach therefore requires technical advances in measurement and mitigation.

Conclusion

LLMs distinguish themselves from previous technologies in their ability to engage in open-ended cognitive processes and act on them, functioning less as tools and more as collaborative thought partners and autonomous agents (Collins et al. 2024b). In this paper, we chart a path to address *overreliance* on LLMs, showing that its consequences range from immediate errors to gradual shifts in human autonomy, skill development, and well-being. We call on the AI community to

treat AI reliance as a central research priority. As LLMs become embedded in all aspects of our daily lives, proactively addressing overreliance will ensure these powerful technologies augment our capabilities instead of inadvertently—or worse, irreversibly—diminishing them.

Appendix

LLM-as-judge Prompts

Each incident’s title and description from the AI Incident Database ($N = 1,339$) were classified by three LLM judges: GPT-4o-mini, GPT-4.1, and Claude Sonnet 4.5 ((claude-sonnet-4-5-20250929)). The system prompt used for classification is presented below. Labeling was completed using temperature = 0 and max_tokens = 50. An incident was labeled as involving overreliance only if *all three* models returned {"overreliance": true} (i.e., unanimous agreement). Of the 1,339 incidents, 147 were labeled as overreliance.

System prompt

```
You are an expert analyst identifying AI
  ↳ incidents that involve "overreliance
  ↳ on AI."

Overreliance on AI means situations where:
- Humans trusted AI systems too much
  ↳ without adequate verification or
  ↳ oversight
- Users blindly followed AI recommendations
  ↳ without critical evaluation
- Lack of human supervision led to harm
  ↳ when AI made errors
- People deferred to AI judgment when they
  ↳ should have intervened
- Automation complacency - humans became
  ↳ inattentive because they trusted the
  ↳ automation
- Users did not question or verify AI
  ↳ outputs that turned out to be wrong
- AI was used as sole decision-maker when
  ↳ human judgment was needed

This does NOT include:
- AI systems that simply made errors
  ↳ (without human overreliance
  ↳ component)
- Algorithmic bias issues (unless humans
  ↳ failed to verify biased outputs)
- Privacy violations
- AI misuse/malicious use (scams, fraud,
  ↳ deepfakes used to deceive victims)
- Deepfake scams where victims were tricked
  ↳ by AI-generated content they didn't
  ↳ know was AI
- Technical failures without human
  ↳ oversight element
- Cases where someone was deceived BY AI
  ↳ (e.g. voice cloning scams) rather
  ↳ than overrelying ON AI

Respond with ONLY a JSON object:
{"overreliance": true} or {"overreliance":
  ↳ false}
```

Request prompt

Analyze this AI incident for overreliance:

Title: {title}

Description: {description}

Is this an example of overreliance on AI?

Validation

To validate the LLM judge, we drew a stratified random sample of 100 incidents (50 labeled as overreliance, 50 not), shuffled, and hand-coded each by one author who was blind to the LLM labels. The results are in Table 4. We find a mod-

Sample size	100
Accuracy	0.780
Cohen’s κ	0.528
Precision	0.703
Recall	0.703
F1	0.703
Human positive rate	37.0%
LLM ensemble positive rate	37.0%

Table 4: Agreement between the unanimous LLM label and a hand-coded human label on a stratified random sample of 100 incidents (50 drawn from the consensus-positive class and 50 from the consensus-negative class). κ corresponds to “moderate” agreement.

erate level of agreement, which supports the use of these LLM judge labels for a descriptive analysis of trends, similar to what we report here. Making stronger quantitative claims would benefit from additional validation (e.g., using multiple human raters and a larger validation sample).

References

Abdulhai, M.; White, I.; Wan, Y.; Qureshi, I.; Leibo, J.; Kleiman-Weiner, M.; and Jaques, N. 2026. How llms distort our written language. *arXiv preprint arXiv:2603.18161*.

Abercrombie, G.; Curry, A. C.; Dinkar, T.; Rieser, V.; and Talat, Z. 2023. Mirages: On anthropomorphism in dialogue systems. *arXiv preprint arXiv:2305.09800*.

AIID. 2024. Incident 838: Microsoft Copilot Allegedly Provides Unsafe Medical Advice with High Risk of Severe Harm — incidentdatabase.ai. <https://incidentdatabase.ai/cite/838/>. [Accessed 09-05-2025].

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–13.

Ashkinaze, J.; Mendelsohn, J.; Li, Q.; Budak, C.; and Gilbert, E. 2024. How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment. *arXiv preprint arXiv:2401.13481*.

Azzutti, A. 2024. Artificial Intelligence and Machine Learning in Finance: Key Concepts, Applications, and Regulatory Considerations. In *The Emerald Handbook of Fintech: Reshaping Finance*, 315–339. Emerald Publishing Limited.

Bainbridge, L. 1983. Ironies of automation. In *Analysis, design and evaluation of man-machine systems*, 129–135. Elsevier.

Balayn, A.; Yurrita, M.; Rancourt, F.; Casati, F.; and Gadi-
raju, U. 2024. An Empirical Exploration of Trust Dynamics in LLM Supply Chains. *arXiv preprint arXiv:2405.16310*.

Banker, S.; and Khetani, S. 2019. Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, 38(4): 500–515.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–16.

Barker, M.; Kallina, E.; Ashok, D.; Collins, K.; Casovan, A.; Weller, A.; Talwalkar, A.; Chen, V.; and Bhatt, U. 2023. FeedbackLogs: Recording and incorporating stakeholder feedback into machine learning pipelines. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–15.

Bhatt, U.; Chen, V.; Collins, K. M.; Kamalaruban, P.; Kallina, E.; Weller, A.; and Talwalkar, A. 2025. Learning personalized decision support policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14203–14211.

Bhatt, U.; and Sargeant, H. 2024. When Should Algorithms Resign? A Proposal for AI Governance. *Computer*, 57(10): 99–103.

Bo, J. Y.; Wan, S.; and Anderson, A. 2024. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. *arXiv preprint arXiv:2412.15584*.

Bonaccio, S.; and Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2): 127–151.

boyd, d. 2025. We need an interventionist mindset.

Brinkmann, L.; Baumann, F.; Bonnefon, J.-F.; Derex, M.; Müller, T. F.; Nussberger, A.-M.; Czaplicka, A.; Acerbi, A.; Griffiths, T. L.; Henrich, J.; et al. 2023. Machine culture. *Nature Human Behaviour*, 7(11): 1855–1868.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1): 1–21.

Budzyń, K.; Romańczyk, M.; Kitala, D.; Kołodziej, P.; Bugajski, M.; Adami, H. O.; Blom, J.; Buszkiewicz, M.; Halvorsen, N.; Hassan, C.; et al. 2025. Endoscopist deskill risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10): 896–903.

- Cacioppo, J. T.; and Petty, R. E. 1982. The need for cognition. *Journal of personality and social psychology*, 42(1): 116.
- Cao, S.; and Huang, C.-M. 2022. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–23.
- Casper, S.; Krueger, D.; and Hadfield-Menell, D. 2025. Pitfalls of Evidence-Based AI Policy. *arXiv preprint arXiv:2502.09618*.
- Castel, P. K. 2023. *Mata v. Avianca, Inc.* United States District Court, Southern District of New York. No. 1:2022cv01461, Document 54 (S.D.N.Y. 2023).
- Chan, A.; Salganik, R.; Markelius, A.; Pang, C.; Rajkumar, N.; Krashennikov, D.; Langosco, L.; He, Z.; Duan, Y.; Carroll, M.; et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 651–666.
- Chater, N.; Zhu, J.-Q.; Spicer, J.; Sundh, J.; León-Villagrà, P.; and Sanborn, A. 2020. Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, 29(5): 506–512.
- Chatterji, A.; Cunningham, T.; Deming, D. J.; Hitzig, Z.; Ong, C.; Shan, C. Y.; and Wadman, K. 2025. How people use ChatGPT. Technical report, National Bureau of Economic Research.
- Chayka, K. 2025. *Filterworld: How algorithms flattened culture*. Random House.
- Chen, A.; Kim, S. S. Y.; Dharmasiri, A.; Russakovsky, O.; and Fan, J. E. 2025. Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713958.
- Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2): 1–32.
- Cheng, M.; Yu, S.; Lee, C.; Khadpe, P.; Ibrahim, L.; and Jurafsky, D. 2025. Social Sycophancy: A Broader Understanding of LLM Sycophancy. *arXiv preprint arXiv:2505.13995*.
- Chevi, R.; Inui, K.; Solorio, T.; and Aji, A. F. 2025. How Individual Traits and Language Styles Shape Preferences In Open-ended User-LLM Interaction: A Preliminary Study. *arXiv preprint arXiv:2504.17083*.
- Choudhury, A.; and Chaudhry, Z. 2024. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*, 26: e56764.
- Collins, K. M.; Bhatt, U.; and Sucholutsky, I. 2025. Revisiting Rogers' Paradox in the Context of Human-AI Interaction. *arXiv preprint arXiv:2501.10476*.
- Collins, K. M.; Chen, V.; Sucholutsky, I.; Kirk, H. R.; Sadek, M.; Sargeant, H.; Talwalkar, A.; Weller, A.; and Bhatt, U. 2024a. Modulating language model experiences through frictions. *arXiv preprint arXiv:2407.12804*.
- Collins, K. M.; Sucholutsky, I.; Bhatt, U.; Chandra, K.; Wong, L.; Lee, M.; Zhang, C. E.; Zhi-Xuan, T.; Ho, M.; Mansinghka, V.; et al. 2024b. Building machines that learn and think with people. *Nature human behaviour*, 8(10): 1851–1863.
- Csapo, G.; Kim, J.; Klasinc, M.; and ElKattan, A. 2019. Survival of the best fit. USA. <https://www.survivalofthebestfit.com>.
- Csernaton, R. 2024. Can Democracy Survive the Disruptive Power of AI? — [carnegieendowment.org](https://carnegieendowment.org/research/2024/12/can-democracy-survive-the-disruptive-power-of-ai?lang=en). <https://carnegieendowment.org/research/2024/12/can-democracy-survive-the-disruptive-power-of-ai?lang=en>. [Accessed 09-05-2025].
- Cui, R.; Li, M.; and Zhang, S. 2022. AI and procurement. *Manufacturing & Service Operations Management*, 24(2): 691–706.
- Cummings, M. L. 2006. Automation and accountability in decision support system interface design.
- Deterding, S.; Hook, J.; Fiebrink, R.; Gillies, M.; Gow, J.; Akten, M.; Smith, G.; Liapis, A.; and Compton, K. 2017. Mixed-initiative creative interfaces. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 628–635.
- Detommaso, G.; Bertran, M.; Fogliato, R.; and Roth, A. 2024. Multicalibration for confidence scoring in llms. *arXiv preprint arXiv:2404.04689*.
- Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1): 114.
- Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3): 1155–1170.
- Dzindolet, M. T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; and Beck, H. P. 2003. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6): 697–718.
- Earp, B. D.; Mann, S. P.; Aboy, M.; Awad, E.; Betzler, M.; Botes, M.; Calcott, R.; Caraccio, M.; Chater, N.; Coeckelbergh, M.; et al. 2025. Relational norms for human-AI co-operation. *arXiv preprint arXiv:2502.12102*.
- Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; et al. 2025. How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*.
- Feng, K. K.; Pu, K.; Latzke, M.; August, T.; Siangliu-lue, P.; Bragg, J.; Weld, D. S.; Zhang, A. X.; and Chang, J. C. 2024. Cocoa: Co-Planning and Co-Execution with AI Agents. *arXiv preprint arXiv:2412.10999*.

- Gerlich, M. 2025. The human factor of AI: Implications for critical thinking and societal anxieties. *Available at SSRN* 5085897.
- Glikson, E.; and Woolley, A. W. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, 14(2): 627–660.
- Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.
- Griffiths, T. L. 2020. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11): 873–883.
- Griffiths, T. L.; Chater, N.; and Tenenbaum, J. B. 2024. *Bayesian models of cognition: reverse engineering the mind*. MIT Press.
- Grunde-McLaughlin, M.; Mozannar, H.; Murad, M.; Chen, J.; Amershi, S.; and Fourney, A. 2026. Overseeing Agents Without Constant Oversight: Challenges and Opportunities. *arXiv preprint arXiv:2602.16844*.
- Guo, Z.; Wu, Y.; Hartline, J. D.; and Hullman, J. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 221–236.
- Harvey, N.; and Fischer, I. 1997. Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes*, 70(2): 117–133.
- He, G.; Demartini, G.; and Gadiraju, U. 2025. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–22.
- He, G.; Kuiper, L.; and Gadiraju, U. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- Hoff, K. A.; and Bashir, M. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3): 407–434.
- Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159–166.
- Hou, Y. T.-Y.; and Jung, M. F. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.
- Huang, Y.; Gao, C.; Zhou, Y.; Guo, K.; Wang, X.; Cohen-Sasson, O.; Lamparth, M.; and Zhang, X. 2025. Position: We Need An Adaptive Interpretation of Helpful, Honest, and Harmless Principles. *arXiv preprint arXiv:2502.06059*.
- Hullman, J.; Kale, A.; and Hartline, J. 2025. Decision theoretic foundations for experiments evaluating human decisions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15.
- Hunter, R.; Moulange, R.; Bernardi, J.; and Stein, M. 2024. Monitoring Human Dependence On AI Systems With Reliance Drills. *arXiv preprint arXiv:2409.14055*.
- Ibrahim, L.; Akbulut, C.; Elasmr, R.; Rastogi, C.; Kahng, M.; Morris, M. R.; McKee, K. R.; Rieser, V.; Shanahan, M.; and Weidinger, L. 2025. Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models. *arXiv preprint arXiv:2502.07077*.
- Ibrahim, L.; Ghassemi, M. M.; and Alhanai, T. 2023. Do explanations improve the quality of AI-assisted human decisions? An algorithm-in-the-loop analysis of factual & counterfactual explanations. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 326–334.
- Ibrahim, L.; Hafner, F. S.; Cheng, M.; Lee, C.; Anselmetti, R.; Willer, R.; Rocher, L.; and Yang, D. 2026. Sycophantic AI makes human interaction feel more effortful and less satisfying over time. *arXiv preprint arXiv:2605.07912*.
- Ibrahim, L.; Hafner, F. S.; and Rocher, L. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*.
- Ibrahim, L.; Huang, S.; Bhatt, U.; Ahmad, L.; and Anderljung, M. 2024. Towards interactive evaluations for interaction harms in human-AI systems. *arXiv preprint arXiv:2405.10632*.
- Kahr, P. K.; Rooks, G.; Snijders, C.; and Willemsen, M. C. 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, 609–622. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705083.
- Kelly, M.; Kumar, A.; Smyth, P.; and Steyvers, M. 2023. Capturing Humans' mental models of AI: An item response theory approach. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 1723–1734.
- Kim, S. S. Y.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–835.
- Kim, S. S. Y.; Vaughan, J. W.; Liao, Q. V.; Lombrozo, T.; and Russakovsky, O. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *ACM Conference on Human Factors in Computing Systems (CHI)*.
- Kim, S. S. Y.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 77–88. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.

- Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.
- Kudina, O.; and de Boer, B. 2024. Large language models, politics, and the functionalization of language. *AI and Ethics*, 1–13.
- Lai, V.; Chen, C.; Smith-Renner, A.; Liao, Q. V.; and Tan, C. 2023a. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369–1385.
- Lai, V.; Zhang, Y.; Chen, C.; Liao, Q. V.; and Tan, C. 2023b. Selective explanations: Leveraging human input to align explainable AI. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–35.
- Lee, H.-P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; and Wilson, N. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, 1–22.
- Lee, J. D.; and Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10): 1243–1270.
- Lee, J. D.; and Moray, N. 1994. Trust, self-confidence, and operators’ adaptation to automation. *International journal of human-computer studies*, 40(1): 153–184.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1): 50–80.
- Lieder, F.; and Griffiths, T. L. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43: e1.
- Liu, R.; Geng, J.; Peterson, J. C.; Sucholutsky, I.; and Griffiths, T. L. 2024. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.
- Logg, J. M.; Minson, J. A.; and Moore, D. A. 2019. AI-algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151: 90–103.
- Long, D.; and Magerko, B. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.
- Lu, Y.; Zhang, C.; Yang, Y.; Yao, Y.; and Li, T. J.-J. 2024. From awareness to action: exploring end-user empowerment interventions for dark patterns in UX. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–41.
- Lu, Z.; and Yin, M. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.
- Ma, S.; Lei, Y.; Wang, X.; Zheng, C.; Shi, C.; Yin, M.; and Ma, X. 2023. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Manzey, D.; Reichenbach, J.; and Onnasch, L. 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1): 57–87.
- Manzini, A.; Keeling, G.; Alberts, L.; Vallor, S.; Morris, M. R.; and Gabriel, I. 2024. The Code That Binds Us: Navigating the Appropriateness of Human-AI Assistant Relationships. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 943–957.
- McKee, K. R.; Bai, X.; and Fiske, S. T. 2024. Warmth and competence in human-agent cooperation. *Autonomous Agents and Multi-Agent Systems*, 38(1): 23.
- Messeri, L.; and Crockett, M. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002): 49–58.
- Metzger, L.; Miller, L.; Baumann, M.; and Kraus, J. 2024. Empowering calibrated (dis-) trust in conversational agents: a user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Naiseh, M.; Al-Mansoori, R. S.; Al-Thani, D.; Jiang, N.; and Ali, R. 2021. Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, 1–5. IEEE.
- Nishal, S.; Sinchai, J.; and Diakopoulos, N. 2024. Understanding practices around computational news discovery tools in the domain of science journalism. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–36.
- Oktar, K.; Collins, K. M.; Hernandez-Orallo, J.; Coyle, D.; Cave, S.; Weller, A.; and Sucholutsky, I. 2025. Identifying, Evaluating, and Mitigating Risks of AI Thought Partnerships. *arXiv:2505.16899*.
- O’Neill, O. 2018. Linking trust to trustworthiness. *International Journal of Philosophical Studies*, 26(2): 293–300.
- Padmakumar, V.; Ibrahim, L.; Wang, Z. Z.; Wang, J.; Liao, Q. V.; and Yang, D. 2026. Offloading Score: Measuring AI Reliance Through Counterfactual Workflows. *arXiv preprint arXiv:2605.29392*.
- Parasuraman, R.; and Manzey, D. H. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3): 381–410.
- Parasuraman, R.; and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2): 230–253.
- Passi, S.; Dhanorkar, S.; and Vorvoreanu, M. 2024. Appropriate reliance on Generative AI: Research synthesis. Technical Report MSR-TR-2024-7, Microsoft.

- Passi, S.; and Vorvoreanu, M. 2022. Overreliance on AI: Literature Review. Technical Report MSR-TR-2022-12, Microsoft.
- Pataranutaporn, P.; Liu, R.; Finn, E.; and Maes, P. 2023. Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10): 1076–1086.
- Phang, J.; Lampe, M.; Ahmad, L.; Agarwal, S.; Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W.; Pataranutaporn, P.; et al. 2025. Investigating Affective Use and Emotional Well-being on ChatGPT. *arXiv preprint arXiv:2504.03888*.
- Pu, K.; Lazaro, D.; Arawjo, I.; Xia, H.; Xiao, Z.; Grossman, T.; and Chen, Y. 2025. Assistance or Disruption? Exploring and Evaluating the Design and Trade-offs of Proactive AI Programming Support. *arXiv preprint arXiv:2502.18658*.
- Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW1): 1–22.
- Robrecht, A. S.; Rothgänger, M.; and Kopp, S. 2023. A Study on the Benefits and Drawbacks of Adaptivity in AI-generated Explanations. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 1–8.
- Sarkar, A. 2024. Intention Is All You Need. *arXiv preprint arXiv:2410.18851*.
- Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; and Satzger, G. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 410–422.
- Sharma, M.; McCain, M.; Douglas, R.; and Duvenaud, D. 2026. Who’s in Charge? Disempowerment Patterns in Real-World LLM Usage. *arXiv preprint arXiv:2601.19062*.
- Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Shorinwa, O.; Mei, Z.; Lidard, J.; Ren, A. Z.; and Majumdar, A. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*.
- Si, C.; Goyal, N.; Wu, S. T.; Zhao, C.; Feng, S.; Daumé III, H.; and Boyd-Graber, J. 2023. Large Language Models Help Humans Verify Truthfulness—Except When They Are Convincingly Wrong. *arXiv preprint arXiv:2310.12558*.
- Simkute, A.; Tankelevitch, L.; Kewenig, V.; Scott, A. E.; Sellen, A.; and Rintel, S. 2025. Ironies of generative AI: understanding and mitigating productivity loss in Human-AI interaction. *International Journal of Human-Computer Interaction*, 41(5): 2898–2919.
- Sniezek, J. A.; and Van Swol, L. M. 2001. Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes*, 84(2): 288–307.
- Spatharioti, S. E.; Rothschild, D.; Goldstein, D. G.; and Hoffman, J. 2025. Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In *CHI 2025*.
- Steyvers, M.; and Kumar, A. 2024. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, 19(5): 722–734.
- Steyvers, M.; Tejada, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.
- Subramonyam, H.; Pea, R.; Pondoc, C.; Agrawala, M.; and Seifert, C. 2024. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022): 1279–1285.
- Tennant, E.; Henke, B.; Keshmirian, A.; Shanahan, M.; Rieser, V.; Lum, K.; Levine, S.; and Haas, J. 2026. Normative Robustness as a Frontier for Non-Verifiable Reasoning in LLMs. *arXiv preprint arXiv:2606.12731*.
- Turchi, T.; Malizia, A.; Paternò, F.; Borsci, S.; and Chamberlain, A. 2024. Adaptive XAI: Towards intelligent interfaces for tailored AI explanations. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, 119–121.
- Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *science*, 185(4157): 1124–1131.
- Vafa, K.; Rambachan, A.; and Mullainathan, S. 2024. Do large language models perform the way people expect? measuring the human generalization function. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Van Dongen, K.; and Van Maanen, P.-P. 2013. A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, 71(4): 410–424.
- Vasconcelos, H.; Bansal, G.; Fournay, A.; Liao, Q. V.; and Wortman Vaughan, J. 2025. Generation probabilities are not enough: Uncertainty highlighting in ai code completions. *ACM Transactions on Computer-Human Interaction*, 32(1): 1–30.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–38.
- Vereschak, O.; Bailly, G.; and Caramiaux, B. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–39.
- Vodrahalli, K.; Gerstenberg, T.; and Zou, J. Y. 2022. Uncalibrated models can improve human-AI collaboration. *Advances in Neural Information Processing Systems*, 35: 4004–4016.

- Wang, X.; Lu, Z.; and Yin, M. 2022. Will you accept the AI recommendation? predicting human behavior in AI-assisted decision making. In *Proceedings of the ACM web conference 2022*, 1697–1708.
- Wang, X.; and Yin, M. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 318–328.
- Weisz, J. D.; He, J.; Muller, M.; Hoefer, G.; Miles, R.; and Geyer, W. 2024. Design principles for generative AI applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22.
- Yin, M.; Vaughan, J. W.; and Wallach, H. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems*.
- Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8653–8665. Toronto, Canada: Association for Computational Linguistics.
- Ying, L.; Collins, K. M.; Wong, L.; Sucholutsky, I.; Liu, R.; Weller, A.; Shu, T.; Griffiths, T. L.; and Tenenbaum, J. B. 2025. On Benchmarking Human-Like Intelligence in Machines. *arXiv preprint arXiv:2502.20502*.
- Yu, S.; Cheng, M.; Jabbar, A.; Sucholutsky, I.; Collins, K. M.; Jurafsky, D.; and Hawkins, R. D. 2026. The efficiency-gain illusion: People underestimate the rate of AI use and overestimate its benefits on simple tasks. *arXiv preprint arXiv:2605.22687*.
- Zao-Sanders, M. 2025a. How people are really using Gen AI in 2025. *Harvard Business Review*. Accessed: 2025-07-15.
- Zao-Sanders, M. 2025b. How People Are Really Using Gen AI in 2025. *Harvard Business Review*.
- Zerilli, J.; Bhatt, U.; and Weller, A. 2022. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4).
- Zhai, C.; Wibowo, S.; and Li, L. D. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1): 28.
- Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305.
- Zhou, K.; Hwang, J. D.; Ren, X.; and Sap, M. 2024. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. *arXiv preprint arXiv:2401.06730*.