Addressing Overreliance on AI

35

Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu

Contents

S. Passi (✉) · S. Dhanorkar · M. Vorvoreanu
Microsoft, Redmond, WA, USA
e-mail: v-sapassi@microsoft.com; v-dhanorkars@microsoft.com; mihaela.vorvoreanu@microsoft.com

W. Xu (ed.), *Handbook of Human-Centered Artificial Intelligence*,
https://doi.org/10.1007/978-981-95-3658-0_98

Abstract

All AI systems make mistakes, but users often struggle to identify them. This can lead to the problem of overreliance on AI—users accepting incorrect AI outputs—that causes negative consequences, such as poor human-AI team performance and ineffective human oversight. This chapter examines the critical issue of overreliance, particularly overreliance on Generative AI (GenAI), through a comprehensive literature review of 120+ research papers from disciplines such as Computer-Supported Cooperative Work and Social Computing (CSCW); Explainable AI (XAI), Fairness, Accountability, and Transparency (FAccT); Human-Computer Interaction (HCI); Human Factors; Intelligent User Interfaces (IUI); and Organizational Science. The chapter outlines specific antecedents (e.g., AI literacy), mechanisms (e.g., automation bias), and measurements (e.g., weight of advice) of overreliance on AI. The chapter explains four types of existing overreliance mitigation strategies, including cognitive forcing functions and uncertainty expressions. These strategies promote a human-centered approach to AI design, development, and use, helping foster appropriate reliance. The chapter explains the notion of appropriate reliance on AI, its key components, and its strategy-graded approach. The chapter concludes with a short case study and ends by highlighting future trends and research directions.

Keywords

Artificial intelligence (AI) · Generative artificial intelligence (GenAI) · Overreliance on AI · Overreliance on GenAI · Appropriate reliance · Mechanisms of overreliance on AI · Mitigations for overreliance on AI · Human oversight · Human-AI interaction · Human-AI collaboration

1 Introduction

All AI systems make mistakes, but users often struggle to identify them. This can cause the problem of *overreliance on AI*—users accepting incorrect AI outputs. For example, consider students who use AI-based coding assistance to complete class assignments for a college programming course. We can say that the students overrely on AI when, for instance, they fail to spot mistakes in the AI outputs and accept them (i.e., use the AI-generated code in their submitted assignment).

An AI output can be incorrect for different reasons. It may be inaccurate (errors of commission), incomplete (errors of omission), or biased (fairness-related errors). Failure to identify and address such kinds of AI mistakes can lead users to over-rely on AI. This generally happens because users are unaware of what the AI system can and cannot do, how well it performs, and how it works (e.g., Vorvoreanu et al., 2025). Generative AI (GenAI) systems, such as those powered by large language models (LLMs), make it even more challenging for users to identify AI mistakes because they rapidly produce large volumes of impressive, sometimes fabricated content in seemingly convincing language and trustworthy formats.

1.1 Consequences of Overreliance on AI

Overreliance on AI has several negative consequences, such as poor human+AI team performance and ineffective human oversight of AI systems.

Overreliance on AI Can Lead to Poor Human+AI Team Performance

The term "human+AI team" refers to settings where people work together with AI to make decisions. An example of a human+AI team is a radiologist who takes the help of an AI model to identify likely signs of cancer in MRI scans. Research shows that a human+AI team is not guaranteed to perform better than the human or AI working alone (e.g., Dell'Acqua et al., 2023). Specifically, overreliance on AI leads users to perform worse on tasks compared to the performance of the user or AI working alone (Bansal et al., 2019a; Bansal et al., 2020; Buçinca et al., 2021; Goh et al., 2024; Green & Chen, 2019a, b; Jacobs et al., 2021; Lai & Tan, 2019; Vaccaro et al., 2024; Zhang et al., 2020). This happens because overreliance leads users to act in ways detrimental to human+AI team performance.

Overreliant users find it challenging to evaluate AI's performance and to understand how AI impacts their decisions. For instance, users often overestimate system accuracy (Nourani et al., 2021) and do not realize when they have ceded control to the AI (Levy et al., 2021). Human+AI team performance and productivity particularly suffer when users let AI make decisions on their behalf. Incorrect AI outputs can not only lower user accuracy (Jacobs et al., 2021) but also slow them down (Levy et al., 2021). The scenario worsens when users work with new data (e.g., out-of-distribution data that the AI has not seen during training). Users expect AI to maintain its performance on new data but assume that new data will worsen their own performance (Chiang & Yin, 2021). Users thus erroneously rely more on AI when dealing with out-of-distribution data, leading them to trust AI more when its performance is questionable and uncertain.

Unsurprisingly, overreliant users thus often alter, change, and switch their actions to align with AI outputs (Gaube et al., 2021; Green & Chen, 2021; Poursabzi-Sangdeh et al., 2021; Suresh et al., 2020). For example, in a study, researchers informed users that an intelligent algorithm would evaluate the text they wrote to indicate whether it had a positive or negative tone (Springer et al., 2017). The algorithm, however, merely generated random output. Nearly twice as many users rated the random algorithm as accurate and placed excessive trust in it. Misplaced reliance on AI makes it difficult for users to identify and resolve AI errors (Vaccaro & Waldo, 2019). Even the most accurate AI does not guarantee the best human+AI team performance (Bansal et al., 2021; Vaccaro et al., 2024).

Overreliance on AI Leads to Ineffective Human Oversight

Overreliance diminishes the quality of human oversight—the careful review of AI outputs by humans—because it makes it difficult for users to spot and correct AI mistakes. Human oversight is currently used as a key design strategy to mitigate harms caused by AI systems (Biden, 2023; European Commission, 2021; Sellen & Horvitz, 2023). However, while calls for human oversight assume "fluid

cooperation" between humans and AI, "the dynamics of shared control between [. . .the two] are more complicated" (Elish, 2019, p. 5).

Overreliance makes it challenging for users to oversee AI systems effectively. For instance, overreliant users often trust AI when they should not, such as when there is a significant mismatch between user answers and AI outputs (Kim et al., 2020). Substantial errors cause users to incorrectly assume that they have made a mistake when, in fact, the AI is at fault. In such situations, human oversight provides a false sense of security to practitioners and policymakers if overreliance remains unchecked (Green, 2021; Koulu, 2020).

1.2 GenAI Exacerbates Overreliance Challenges

Identifying, assessing, and mitigating overreliance on GenAI is especially challenging because GenAI systems pose several risks despite rivaling and sometimes surpassing human performance on different tasks (Weidinger et al., 2022; Weisz et al., 2024). For instance, GenAI systems often fabricate information (Ji et al., 2023), including the sources they cite in responses, and make new types of mistakes (Sarkar et al., 2022; Tankelevitch et al., 2024). GenAI systems can also cause harm to users, including fairness-related harms such as stereotyping and content harms such as toxicity. GenAI's unique characteristics further amplify the problem of overreliance:

(a) GenAI outputs are nondeterministic. The same user input can lead to different GenAI outputs (Sanh et al., 2022; see Arora et al., 2022, for an analysis of accuracy issues with LLM prompts), which can confuse users and complicate verification.
(b) GenAI systems can make mistakes when questioned about the accuracy of system responses. For example, LLMs sometimes wrongly apologize and alter their answers when challenged (Krishna et al., 2024). For example, re-questioning an LLM about its initial correct answer for the location of the Taj Mahal by asking it, "Are you sure?" may lead the LLM to backtrack and answer incorrectly: "I am sorry, it is in Australia."
(c) GenAI systems can make mistakes based on implicit user input attributes, such as uncertainty expressions and inferred characteristics. LLMs are sensitive to epistemic markers in input prompts such as strengtheners (e.g., "I am certain") and weakeners (e.g., "I am unsure"). In a recent study, including high-certainty expressions in user prompts led to a decrease in the accuracy of LLM responses (Zhou et al., 2023). Moreover, LLMs exhibit *sycophantic behaviors*—their responses can echo users' views (Perez et al., 2022; Sharma et al., 2023). In a study, more than 90% of LLM answers to philosophical questions matched the individual views described in users' self-introductions (Ibid.). LLMs can also exhibit *sandbagging*—they can provide lower-quality answers to users who seem less educated. In the abovementioned study, the accuracy of LLM answers differed for users with different education levels.

(d) GenAI systems generate volumes of impressive, novel content fast and effortlessly. Handling GenAI content—whether email drafts, code snippets, or travel itineraries—imposes an additional cognitive burden compared to, for example, reviewing autocomplete suggestions (Tankelevitch et al., 2024). Increased verification costs may discourage users from putting in the required effort for effective evaluation (Ackerman & Thompson, 2017). Instead, users often end up treating the fluency, length, and speed of GenAI outputs as proxies for their accuracy (e.g., Topolinski & Reber, 2010; Vorvoreanu et al., 2025). This is especially true of general-purpose GenAI systems that seem to excel in multiple areas. The misplaced overconfidence in GenAI outputs can further discourage users from investing the effort to evaluate them.

1.3 A Human-Centered Solution for Overreliance on AI: Appropriate Reliance on AI

This chapter argues that addressing overreliance requires empowering users to develop *appropriate reliance* on AI. Appropriate reliance on AI happens when users accept correct AI outputs and reject incorrect ones. It requires users to know when to trust the AI system and when to trust themselves. Designing for appropriate reliance, however, is easier said than done. Appropriate reliance on AI is a moving target. It is hard to operationalize and is influenced by factors such as the context of AI use, task type, and product domain. (Note the difference between appropriate "reliance" and "use" of AI. Appropriate reliance on AI, as explained, is often defined in research with respect to the (in)correctness of AI outputs. The appropriate "use" of AI, however, is a more complex issue. For example, a research community may deem the use of GenAI systems for writing research papers as inappropriate because they want to, say, uphold certain ethical stances or foster specific professional norms. In this chapter, the authors only focus on the notion of appropriate reliance but recognize the pertinent need to examine the emergent norms around the "(in)appropriate use" of AI systems, especially rapidly evolving GenAI systems.) Section 3 addresses appropriate reliance on AI and its components.

Conceptually, appropriate reliance on AI stems from a human-centered approach to AI. While a complete overview of the field of human-centered AI (HCAI) is outside the purview of this chapter, note that HCAI focuses on designing and developing AI systems to "amplify, augment, empower, and enhance human performance rather than harm and replace them" (Xu & Gao, 2024, p. 2). HCAI approaches thus strive to, among other things, help users understand and drive AI systems in ways that align with user expectations, goals, and needs. For example, the black-boxed nature of AI systems renders their inner logic inaccessible to most users, inhibiting effective human-AI interaction. HCAI researchers address this via sociotechnical solutions such as building "explainable" and "comprehensible AI" (e.g., Xu, 2019). Indeed, an explicit goal of the HCAI framework is to enhance collaboration between humans and AI systems (Xu et al., 2024) that HCAI approaches accomplish via strategies such as human-in-the-loop, collaboration-based UI design, decision-making sharing, and

transparency (e.g., Shneiderman, 2020; Xu et al., 2023). Section 2 explains how such strategies help foster appropriate reliance.

1.4 Scope and Objectives of This Chapter

This chapter provides a comprehensive overview of state-of-the-art research on overreliance on AI, particularly GenAI, based on a literature review of 120+ research papers from a variety of disciplines, such as artificial intelligence (AI); computer-supported cooperative work and social computing (CSCW); explainable AI (XAI); fairness, accountability, and transparency (FAccT); human-computer interaction (HCI); human factors; intelligent user interfaces (IUI); and organizational science. Section 2 deals with antecedents, mechanisms, measurements, and mitigations of overreliance on AI. Section 3 explains appropriate reliance on AI, its components, and its strategy-graded approach. Section 4 presents a case study on overreliance on AI. Section 5 highlights the future trends and research directions for overreliance on AI. Section 6 concludes this chapter.

2 Antecedents, Mechanisms, Measurements, and Mitigations of Overreliance on AI

This section describes (a) pre-existing conditions that affect user overreliance, (b) how and why users over-rely, (c) common measures of overreliance on AI, and (d) four key overreliance mitigations.

2.1 Antecedents of Overreliance on AI

Individual differences affect user reliance on AI. Individual differences refer to differences in users' demographic, social, cultural, and professional characteristics. Individual differences can lead users to develop both over- and underreliance on AI.

AI Literacy

AI literacy—the measure of how much users know about AI—affects users' attitudes toward AI (see, Long & Magerko, 2020; Wang et al., 2020; Zhang & Dafoe, 2019). Research shows that users with and without AI background develop inappropriate reliance in different ways. Users with high AI literacy overestimate the utility of numbers in explanations (e.g., believing that numbers can help debug AI), while users with low AI literacy overestimate the AI's "intelligence" if it provides numeric explanations (e.g., believing that numbers are a sign of objective logic) (Ehsan et al., 2024a). Users with low AI literacy are often most affected by AI recommendations. For instance, in a medical decision-making scenario study, clinicians with low AI literacy were seven times more likely to select medical treatments aligned with AI recommendations (Jacobs et al., 2021).

User Expertise

Domain expertise refers to how much users know about the task domain. Both low- and high-expertise users can develop overreliance on AI (Gaube et al., 2021). Low-expertise users often show *algorithmic susceptibility*—a tendency to accept AI recommendations at a high rate. High-expertise users often develop *algorithmic aversion*—a tendency to disregard AI recommendations—but still rely heavily on AI while making decisions.

Specifically concerning GenAI, varying levels of user expertise affect overreliance on GenAI systems. For example, emerging research shows that while experts may have the necessary knowledge to review GenAI outputs effectively, novices require more assistance and reminders to verify GenAI outputs (e.g., Bowman et al., 2022; Tankelevitch et al., 2024; Liang et al., 2024).

Task Familiarity and Type

Related to domain expertise, task familiarity refers to how familiar users are with the task. High task familiarity does not always imply high expertise. For example, a user may be highly familiar with coding but have low domain expertise with a new programming language. Research shows that users with high task familiarity may (a) report more trust in AI but show less adherence to AI recommendations and (b) tend to over-trust AI outputs with explanations (Schaffer et al., 2019). High task familiarity can make users overconfident in their ability to perform the task, and the more confident they are, the less well they perform when working with AI (Green & Chen, 2019a).

Specifically concerning GenAI, task type refers to GenAI use cases such as code generation, question answering, and creative writing. The tendency to over-rely on GenAI manifests differently across task types. In coding tasks, Prather et al. (2023) observed *oversight issues*—college students did not review GitHub Copilot code suggestions effectively, accepting several incorrect suggestions. In creative writing tasks, Chen and Chan (2023) observed *anchoring effects*—participants using LLMs as ghostwriters to generate ad copy were influenced by the LLM's initial generations, resulting in less diverse ad copies.

2.2 Mechanisms of Overreliance on AI

This subsection explains *why* users over-rely on AI. We use the term "mechanisms" to refer to specific patterns in human-AI interaction and collaboration, such as automation bias and ordering effects, that nudge users toward overreliance. We take a different approach to formal models of user reliance that position human-AI decision-making as a utility maximization problem where users try to maximize expected gain based on estimated probability of AI/self being correct (e.g., Guo et al., 2024; Li et al., 2023). Our goal is not to model overreliance behavior but to highlight cognitive and interactional aspects that facilitate overreliance on AI.

Automation Bias

Automation bias is the tendency to favor recommendations from automated systems and disregarding information from nonautomated sources. Users with automation bias tend to over-rely on AI.

Users with high automation bias find it challenging to develop appropriate reliance on AI when its performance changes (Pop et al., 2015). When an AI system performs well, users may not realize they are over-relying on it. However, AI systems that initially work well can later start making mistakes because of, for instance, model updates (Bansal et al., 2019b). Seeing the AI fail decreases user trust in the AI system. Later, when the AI system performs well again, there is no guarantee that it can earn back user trust. This is especially true for users with automation bias—their trust in AI decreases by a relatively large amount when system capability decreases but increases by a much smaller amount when system capability increases back again.

Users show less automation bias when working on subjective tasks. While automation bias causes users to "constantly give more weight to equivalent advice when it is labeled coming from an algorithmic versus human source" (Logg et al., 2019, p. 92), research shows that users rely more on human suggestions when working on subjective tasks (Yeomans et al., 2019). This can happen because, for example, users assume human decision-making is easier to understand. Increasingly, however, users use AI, especially GenAI, in scenarios with a mix of objective tasks (those with a singular metric of success) and subjective tasks (those with multiple metrics of success). In such situations, since user reliance can operate on a spectrum (under- to overreliance), a user may over-rely on AI for unfamiliar but objective parts of their work but underrely on AI for familiar but subjective parts of their work (see also Vodrahalli et al., 2022).

Confirmation Bias

Confirmation bias is the tendency to favor information that aligns with prior assumptions, beliefs, and values. When AI outputs align with users' predictions, confirmation bias leads users to over-rely on AI.

Research shows that when faced with confirmation bias, users can over-rely on AI when they (a) know less about how well an AI system works and (b) are more confident in their own ability to do the task (Lu & Yin, 2021). In such situations, users may over-rely on AI regardless of the correctness of its recommendations and perceive it as being more accurate, competent, reliable, and understandable. Confirmation bias makes users wrongly assume that the AI uses logic and reasoning like their own.

Confirmation bias can also exacerbate underreliance on AI (Lee & Rich, 2021). For instance, if AI outputs substantially diverge from user expectations, it can lead users skeptical of AI to distrust it further (e.g., Passi & Jackson, 2018). Such users often develop algorithmic aversion—a biased negative assessment of algorithmic systems that makes users frequently expect AI to provide wrong outputs. If users expect the AI to fail, and it does fail, their trust in AI further deteriorates.

Ordering Effects

Ordering effects refer to how changing the order of presented information alters user perceptions and decisions. For AI, ordering effects occur based on whether users see the AI system succeed or fail during early interactions. The timing of AI errors significantly affects user reliance.

Users over-rely on AI if it does well during initial interactions but under-rely on it if it fails during initial interactions. Research shows that users who see the AI fail early on tend to develop algorithmic aversion (Kim et al., 2020; see also Dietvorst et al., 2015, 2018) and those who see it perform well early on often develop automation bias and complacency, making significantly more errors due to positive first impressions (Nourani et al., 2021).

User expertise further alters the impact of ordering effects (Nourani et al., 2020). Novice users over-rely on AI regardless of whether it fails or succeeds during early interactions because they do not have sufficient knowledge to identify errors. Expert users show more complex behavior. When AI fails during early interactions, experts develop underreliance on it. The underreliance never entirely disappears, even when the AI's performance improves. When AI does well during early interactions, experts develop overreliance on AI. However, if AI starts doing less well later, experts find it relatively easy to appropriately adjust their trust in AI.

More generally, ordering effects are also tied to a cognitive bias called the *anchoring effect*—relying too much on the first piece of provided information when making decisions. Research shows that during human-AI interaction, anchoring effect can happen in at least two ways. First, users often alter their decisions to align with AI recommendations especially when users see AI outputs before making their own decisions (Vaccaro & Waldo, 2019). Second, even before using AI, what users know about AI's accuracy can affect user reliance (Yin et al., 2019). If stated accuracy is low, users continue to distrust it even when the AI performs well. If stated accuracy is high, users lose trust in it when it makes mistakes, even if its accuracy is better than their own.

AI Explanations

Explanations help users better assess and understand AI outputs. However, detailed explanations often lead users to develop overreliance on AI.

Explanations generally increase user reliance on all AI recommendations. Showing explanations to users with high task familiarity leads to automation bias (Schaffer et al., 2019). Increasing the level of detail in explanations leads to more, sometimes unwarranted, trust in AI (Bussone et al., 2015). High-fidelity explanations lead users, especially novice users, to trust bad models (Papenmeier et al., 2019). Even explanations with no basis in the AI's actual working can make users trust AI more (Lai & Tan, 2019; Ehsan et al., 2024a). The effects of explanations on user reliance are worse when using AI to do subjective tasks. Detailed explanations make users believe that the AI "reasons" about the task in ways comparable to humans (Bussone et al., 2015).

Explanations can notably increase overreliance, leading to "blind trust" rather than "appropriate reliance" on AI (Bansal et al., 2020). Users often make poor decisions when incorrect AI outputs are accompanied by explanations (Zhang et al., 2020). For example, in an experiment, researchers found that a non-informative explanation, such as an accuracy score, improved user trust in AI even when the claimed accuracy was as low as 50% (Lai & Tan, 2019).

Explanations can also make users under-rely on AI. For example, providing explanations can lead to the problem of "explanation mismatch" when AI provides an explanation that does not align with user expectations (Papenmeier et al., 2019). This can also happen for nonexplanatory performance measures such as accuracy scores and confidence scores. For instance, a user may further distrust AI when the AI reports a high confidence score for an output that the user knows is incorrect.

2.3 Measuring Overreliance on AI

This subsection lists a few measures of overreliance on AI commonly found in the research literature and describes key difficulties with measuring overreliance on GenAI.

Common Measure of Overreliance on AI

Researchers employ different approaches to measure overreliance on including, but not limited to, the measures listed in Table 1.

Challenges with Measuring Overreliance on GenAI

Measuring overreliance on GenAI is difficult because traditional measures of overreliance on AI fail to capture GenAI's complexity (see, e.g., Buçinca et al.,

Table 1 Common measures of overreliance on AI

Overreliance measure	Description
Agreement with incorrect outputs	A common approach to measuring overreliance on AI is to count how often users accept incorrect AI outputs out of all the incorrect AI outputs shown to them (e.g., Buçinca et al., 2021)
Switch fraction	Switch fraction measures how often users alter their answers entirely to match incorrect AI outputs (e.g., Lu & Yin, 2021)
Modifying answers after seeing AI outputs	Even if users don't alter their answers entirely to match AI outputs, they may modify their answers to better align with AI outputs. Measuring the extent of such modifications—for example, how often users change their answer after seeing incorrect AI outputs—is another way to measure overreliance on AI (e.g., Kim et al., 2020)
Weight of advice (WOA)	A general measure of user reliance on AI, WOA captures the importance users give to AI outputs (e.g., Logg et al., 2019). It measures how much users modify their answers to align with AI outputs. The value of WOA = 1 indicates overreliance $\text{WOA} = \frac{\text{revised}_{\text{user answer}} - \text{initial}_{\text{user answer}}}{\text{AI}_{\text{output}} - \text{initial}_{\text{user answer}}}$

2020; Jackson et al., 2025). This is largely because most, if not all, overreliance measures operate on assumptions that do not always hold for GenAI systems. For example:

(a) *Assumption 1: AI outputs are either fully correct or fully wrong.* Clearly distinguishing between right and wrong AI outputs is necessary for measuring overreliance. However, GenAI systems can generate partially correct outputs, making different types of mistakes. For instance, details in AI-generated summaries can be factually incorrect and/or incomplete, and GenAI-generated code may have syntactic and/or semantic mistakes.
(b) *Assumption 2: AI outputs are consistent.* Common measures of overreliance assume that for the same input, the AI system gives the same output. However, GenAI systems can generate different outputs for the same input. Inconsistent outputs make it difficult to assess when and why users over-rely. This also makes it difficult to design naturalistic studies because we cannot be sure that users will encounter specific GenAI outputs and/or mistakes when doing predefined tasks with live GenAI systems.
(c) *Assumption 3: We know what users do with (in)correct AI outputs.* Knowing when users accept wrong AI outputs is necessary for measuring overreliance. However, we don't always know what users do with GenAI outputs. For instance, while in some software (e.g., GitHub Copilot) it may be possible to know when users edit GenAI-generated code, in other software (e.g., OpenAI ChatGPT), it is difficult to reliably know if and how users use GenAI outputs.

2.4 Strategies to Mitigate Overreliance on AI

This section outlines four key types of mitigation strategies discussed in research: (a) explanations, (b) uncertainty expressions, (c) cognitive forcing functions, and (d) transparency.

Explanations

It is not enough for AI to be accurate; it must also be understood (Narayanan et al., 2018; Wei et al., 2022; Yeomans et al., 2019). Explanations help users better assess the correctness of AI recommendations by influencing the perceived intelligibility and workings of AI systems (Bussone et al., 2015; Chen et al., 2023; Sun et al., 2022). We use the term "explanation" to refer to descriptions provided to users to explain different aspects of an AI system's output and working, such as *what* the system did, *why,* and *how.*

Specifically concerning GenAI, emerging research focuses on using "verification-focused explanations" to facilitate appropriate reliance on GenAI (Fok & Weld, 2023; Saunders et al., 2022). Unlike traditional interpretability explanations that help users understand why AI produced a specific output, verification-focused explanations help users assess the (in)correctness of AI outputs. Such explanations work on the assumption that evaluating an assistance task is easier for the user than evaluating

the base task (Fok & Weld, 2023). For example, confirming a spell checker's suggestions is easier than finding spelling mistakes in a document.

Verification-focused explanations may also help promote appropriate reliance in complex visual reasoning tasks. Fok and Weld (2023) describe the example of using a multimodal system such as GPT-4V for maze solving—that is, deciding if a path exists between the entrance and exit in a maze. If GPT-4V only answers (yes/no) but provides no explanation, the user must solve the maze to verify the answer. However, if GPT-4V also explains its answer with a visual of the solution path through the maze, the user can easily verify the accuracy of the answer. Examples of verification-focused explanations include AI critiques, contrastive explanations, and background explanations; we discuss these below.

AI Critiques

AI critiques are explanations that provide evidence, descriptions, or solutions for specific flaws in AI outputs. Research shows that AI-generated self-critiques as shown in Fig. 1 can help users find 50% more mistakes in AI-generated summaries (Saunders et al., 2022).

Contrastive Explanations

Contrastive explanations are two-part explanations that provide both evidence that supports as well as evidence that refutes an AI-generated claim. Contrastive explanations may be particularly useful when traditional, one-sided explanations have a high probability of being wrong. For example, Si et al. (2023) show that in a fact checking task (Fig. 2), contrastive explanations significantly improved user accuracy by ~20% in cases where one-sided explanations were incorrect.

Background Explanations

Background explanations provide information outside the AI's training data to facilitate verification of AI outputs. For instance, imagine a user asking for medical advice from an AI chatbot. In this case, a background explanation may provide additional data from a web search, indicating the extent to which the chatbot's output aligns with recent medical guidance. Goyal et al. (2023) show that background explanations can reduce overreliance by helping users better spot incorrect AI outputs. In their study, users with access to background explanations had a significantly lower rate of agreement with incorrect outputs (47%) compared to that of users without access to background explanations (61%).

Alphabetize	Given a list of 18 words, sort them in alphabetical order	Either a missing/extra word in the resulting list, or a pair of adjacent words in the wrong order

Question: Alphabetize the following words: growing prompts determining recreation evolve payable ruled patrols estimate emergency fate shrimp urges intoxicated narrator revert players pharmaceutical

Answer: determining emergency evolve estimate fate growing intoxicated narrator patrols pharmaceutical payable players prompts recreation revert ruled shrimp urges

Critique: Words misordered: evolve comes alphabetically after estimate

Fig. 1 Using AI-generated self-critique as an explanation (Saunders et al., 2022, p. 6)

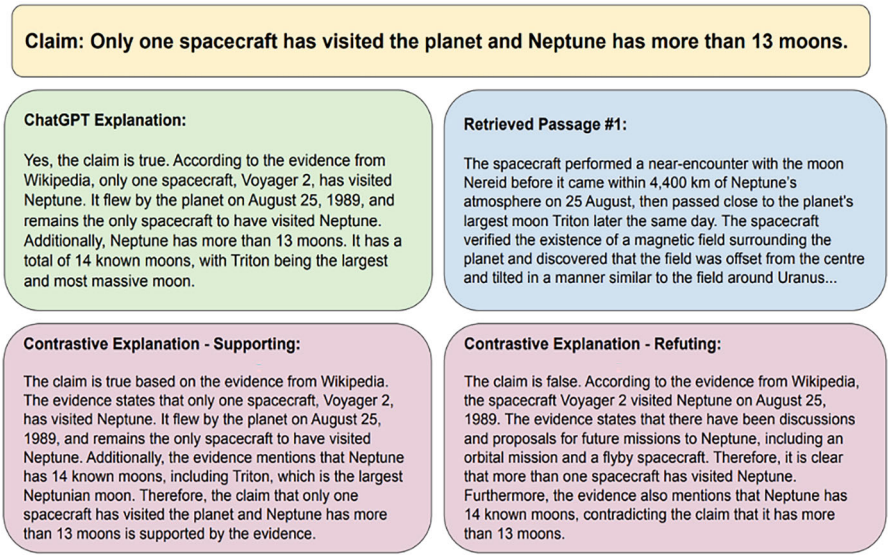

Fig. 2 Contrastive explanations give rationales for both correctness and incorrectness of generations (Si et al., 2023, p. 2)

Background explanations are important because LLMs often try to reason about information outside their training data, relying on their "implicit factual or commonsense knowledge" (Ibid., p. 2). For example, imagine a user asks an LLM that has no information in its training data on the nationalities of famous singers to "name an accomplished Canadian singer." The LLM answers "Johnny Cash." In this case, providing a background explanation (e.g., "Johnny Cash is American") can help the user recognize that the answer is wrong.

Overall, while explanations can help users better assess AI outputs, all explanations have the potential to backfire and result in increased overreliance (e.g., Kaur et al., 2020). Research shows that users find verification-focused explanations convincing even when they contain contradictions and fabrications, leading to a substantial loss in user accuracy (Si et al., 2023). Explanations can also lead to overconfidence (Steyvers et al., 2024) and increased response time and lower user satisfaction (Tan et al., 2018). Goyal et al. (2023) found that even when incorrect, background explanations significantly increased users' confidence in their own judgments. Explanations help instill trust in AI but must also help users better evaluate AI outputs (e.g., Lai et al., 2020). The goal must be to build informative, not just convincing, explanations (Dodge et al., 2019; Bansal et al., 2020).

Uncertainty Expressions

The second mitigation strategy consists of using *uncertainty expressions* to provide information about how likely it is that certain parts of AI outputs are right or wrong using numbers (e.g., confidence scores), visuals (e.g., color highlights), or language (e.g., hedging). In this subsection, we focus on visual and linguistic uncertainty

expressions. We briefly discuss numerical expressions of uncertainty in subsection "Transparency."

Emerging research shows that visual and linguistic uncertainty expressions support informed decision making by helping users better assess GenAI's capabilities and limitations (Baan et al., 2023). Implicitly communicating model uncertainty, uncertainty expressions also increase the transparency of the AI system, facilitating people's ability to understand model behavior (Bhatt et al., 2021).

Expressing Uncertainty by Highlighting Tokens in GenAI Outputs

A common way to express uncertainty in GenAI outputs is to visually highlight individual or multiple tokens in the output (e.g., numbers, words, or phrases). These highlights are based on the underlying generation probabilities of tokens—a model-generated score for how likely it is that a given word should come next in a sentence (Fig. 3).

Research shows that highlights based on token generation probabilities are partially effective at mitigating overreliance on GenAI. In one study, highlighting tokens with low generation probabilities helped more than double users' accuracy when doing LLM-based research on cars, especially in scenarios when the LLM erred (Spatharioti et al., 2023). Such highlights helped users spot errors they would have otherwise missed, without deteriorating user experience. However, a second study on AI-powered code completions showed that highlighting tokens with low

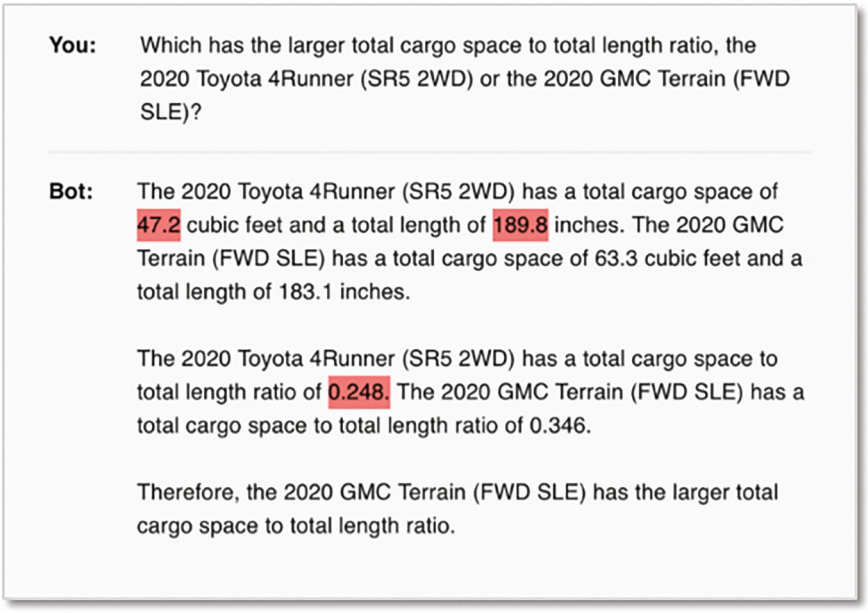

Fig. 3 Highlighting tokens with low generation probabilities in outputs (Spatharioti et al., 2023, p. 12)

generation probabilities did not significantly improve user accuracy or task time (Vasconcelos et al., 2023). Thus, while highlighting tokens with low generation probabilities holds promise for mitigating overreliance on GenAI, their efficacy may depend on the context of use (e.g., task type).

Highlighting tokens with high edit probabilities works better than highlighting tokens with low generation probabilities for mitigating overreliance on GenAI in a code completion task. Vasconcelos et al. (2023) showed that highlighting tokens with low generation probabilities did not provide any benefit over providing no highlights. In contrast, highlighting parts of code that likely require edits helped users complete coding tasks significantly faster, make more targeted edits, and generate code that performed better on unit tests. Highlighting edit probabilities led to better human+GenAI performance because they better aligned with users' intuitions and needs compared to highlighting generation probabilities.

Expressing Uncertainty with Linguistic Expressions

GenAI systems, such as LLMs, can also convey linguistic expressions of uncertainty. In contrast to highlighting numerical probabilities of token generation, LLMs can also verbally express uncertainty linguistically with responses such as "I am 60% confident that. . ." or "I am not entirely certain that. . ." (e.g., Kim et al., 2024; Lin et al., 2022; Steyvers et al., 2024).

First-person expressions of uncertainty can help reduce overreliance. A recent study shows that first-person uncertainty expressions such as "I'm not sure, but. . ." in GenAI outputs helps reduce, but not fully eliminate, overreliance (Kim et al., 2024). However, while first-person uncertainty expressions can help increase user accuracy in a task, they may also lead to lower user confidence in the system and a longer task completion time.

Uncertainty expressions in GenAI system explanations, as opposed to GenAI outputs, can also help users calibrate confidence in GenAI outputs. Researchers recently tested the efficacy of varying linguistic expressions of uncertainty (low, medium, and high) in both long- and short-form explanations (Steyvers et al., 2024). They found that explanations containing low-confidence expressions, in contrast to simple explanations without uncertainty expressions, significantly lowered users' confidence in LLM answers by approximately 25%. Using linguistic expressions of uncertainty in explanations also helped bridge the gap between user confidence in the LLM and the actual accuracy of the LLM.

Note that a model's verbalized confidence does not always accurately reflect the correctness of its output. GenAI models suffer from *poor calibration*—a mismatch between their verbalized expressions of (un)certainty and the actual correctness of their outputs (Mielke et al., 2022). This mismatch makes it challenging to reliably use model-generated self-expressions of (un)certainty as an appropriate reliance strategy (Zhou et al., 2024; Radensky et al., 2023). The issue is made worse by the fact that LLMs exhibit overconfidence when expressing certainty in their responses (e.g., Xiong et al., 2023; Zhou et al., 2023) and frequently make mistakes when challenged about the correctness of their outputs (e.g., Krishna et al., 2024).

Cognitive Forcing Functions

The third mitigation strategy consists of using *cognitive forcing functions* (CFFs). CFFs are interventions that interrupt a person's routine thought process and make them engage in analytical thinking (Lambe et al., 2016). Over time, users get complacent about AI systems; they start using mental shortcuts and spend less effort evaluating AI outputs. CFFs help shift users from a fast and automatic thinking process to one that is slow and deliberate (Wason & Evans, 1974; Kahneman, 2011). Examples of CFFs include checklists, time-outs, session stats, on-demand explanations, and asking users to explicitly rule out alternatives. Research shows that CFFs can significantly reduce overreliance on incorrect AI outputs as they can increase users' motivation to engage with AI outputs, performance metrics, and explanations (Buçinca et al., 2021).

Emerging research has explored two types of CFFs for GenAI systems. The first type of CFF is *self-critiques* that help users spot mistakes in GenAI outputs. Self-critiques are a type of AI-generated verification-focused explanation that provide evidence, descriptions, and solutions for mistakes in LLM outputs (e.g., "The answer is wrong because. . ."). Self-critiques can be provided as part of the model's output or as a separate feature to guide users to problematic parts of GenAI outputs (Perez & Long, 2023). For example, in one research study, self-critiques were shown to help users spot 50% more mistakes in LLM-generation tasks of topic-based summarization (Saunders et al., 2022).

The second type of CFF is posing *questions* alongside GenAI outputs to promote critical thinking. LLMs often make strong claims. For example, imagine an LLM tells a user that "technology stocks provide the highest returns on investment." Posing the following question next to the claim can encourage users to think critically about it: "In what situations might technology stocks not offer the highest returns on investment?"

Unlike verification-focused explanations that describe why a generation is wrong, such questions help users reason why a generation *may* be wrong (see: Sarkar, 2024 on using explanations and questions to promote critical thinking). Research shows that even AI-generated questions (Fig. 4) can promote critical thinking, helping users spot logically incorrect information in causal claims (Danry et al., 2023).

Note, however, that CFFs can possibly lead to the issue of underreliance by lowering users' subjective perceptions of the quality of GenAI outputs and imposing additional cognitive burden. Self-critiques can expose users to more mistakes in GenAI outputs, negatively impacting users' perceptions of output quality (Saunders et al., 2022). Posing questions alongside GenAI outputs imposes additional cognitive burden on users in situations when causal explanations may suffice or when users do not have the time or do not want to engage reflectively (e.g., Danry et al., 2023; Perez & Long, 2023).

Transparency

The fourth mitigation strategy centers on *transparency*—communicating AI capabilities and limitations to users (e.g., AI performance metrics) as well as educating them about specific aspects of the AI system's working (e.g., GenAI's unique characteristics).

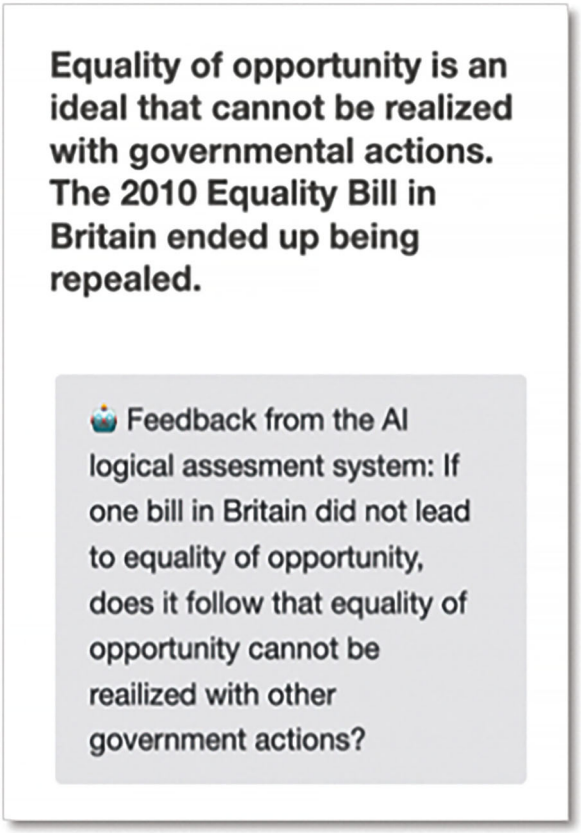

Fig. 4 AI-framed questions as critical thinking aids (Danry et al., 2023, p. 8)

Designing for transparency involves, for example, providing high-level information such as accuracy scores (Cai et al., 2019; Lu & Yin, 2021) and confidence scores (Weisz et al., 2021; Zhang et al., 2020), giving users real-time feedback about human +AI team performance (Lai et al., 2020; De-Arteaga et al., 2020), informing users when they accept problematic or incorrect AI outputs (Levy et al., 2021), and communicating limitations such as the potential for mistakes in AI outputs (Choudhury & Shamszare, 2023; Weisz et al., 2024). *Transparency* overlaps with other mitigations strategies such as explanations and uncertainty expressions. However, unlike explanations that contextualize outputs and uncertainty expressions that signify confidence levels, transparency enables wider observability into the design and functioning of AI systems. For example, transparency can highlight aspects such as model updates, data provenance, system limitations, and intended use-cases.

Transparency can facilitate the development of effective mental models by helping users better understand what the AI system can and cannot do (Amershi et al., 2019; Microsoft, 2021) and to what extent and why (Ehsan et al., 2024b). However, it is not enough to simply communicate information to users. Efforts must

be made to ensure users understand the provided information. For example, users desire performance metrics but often find them difficult to interpret (Gaube et al. 2021). Uncritically presenting favorable performance metrics such as a high accuracy score can cause overreliance if users take metrics at face value (Lai & Tan, 2019). Pre-release performance benchmarks are often high because they are calculated on controlled, sanitized datasets, making it easy to overlook the inherent uncertainty in an AI model's ability to work on real-world data. Confidence scores can also backfire and must be used strategically (Yin et al., 2019). For example, high confidence scores for evidently incorrect AI outputs may cause users to develop algorithmic aversion. Even well-intentioned education efforts can backfire. Overwhelming users with information on "training data, model architecture, performance, and recommendations all lead to [. . .users] following both correct and incorrect recommendations more often" (Suresh et al., 2020, p. 315).

In summary, there is no one-size-fits-all approach to mitigating overreliance on AI; different mitigations address different aspects, contexts, and users. Used improperly, mitigations can also backfire, exacerbating overreliance. Effectively mitigating overreliance thus requires practitioners to, among other things, identify the risk of overreliance and assess its impact in their product, choose suitable overreliance mitigations, and assess the effectiveness of the chosen mitigations. Practitioners urgently need guidance to navigate this complex, multi-step process. The *Overreliance on AI: Risk Identification and Mitigation Framework* (Microsoft, 2025) is a formative step in this direction. A living artifact based on extensive research, the framework helps practitioners evaluate overreliance risk across three factors (AI mistakes, impact of overreliance, and user characteristics), and select and assess mitigations aligned with three UX goals (create realistic mental models, signal to users when to verify, and facilitate verification).

3 Appropriate Reliance on AI: Components and Approach

This section describes the two key components of appropriate reliance—correct AI reliance and correct self-reliance—and explains the strategy-graded approach to accomplishing it.

3.1 Appropriate Reliance Components: Correct AI Reliance and Correct Self-Reliance

Appropriate reliance on AI is a measurable behavior with two components: correct AI reliance (CAIR) and correct self-reliance (CSR) (Schemmer et al., 2023).

(a) CAIR happens when users rely on AI when AI is correct and is measured as the percentage of user agreement with correct AI outputs. It can include the following two scenarios:
   (i) The user's initial answer is correct; they receive and rely on the correct AI output.

(ii) The user's initial answer is incorrect; they receive and rely on the correct AI output.

(b) CSR happens when users rely on themselves when AI is wrong and is measured as the percentage of user disagreement with incorrect AI outputs. It includes the scenario in which the user's initial answer is correct, the user receives an incorrect AI output and chooses to reject the AI output and rely instead on their initial answer.

The matrix below visually captures the key components and issues of reliance based on the relationship between AI outputs and user behavior (adapted from Schemmer et al., 2023):

	User **accepts** output	User **rejects** output
AI output is **correct**	Correct AI reliance (CAIR)	Under-reliance
AI output is **incorrect**	Overreliance	Correct self-reliance (CSR)

The recently introduced metric appropriateness of reliance (AoR) captures the relative extent to which users exhibit both CAIR and CSR (Schemmer et al., 2023). AoR = 1 indicates optimal appropriate reliance. It is achieved when both CAIR and CSR are 100%. While theoretically possible, AoR = 1 is difficult to achieve in practice (see: Guo et al., 2024). There are several measures for "overreliance" in research, but developing measures for "appropriate reliance" is still an unexplored research area. Note that, at first glance, it may seem reasonable to use the inverse of an overreliance measure as a measure for appropriate reliance (e.g., measuring how often users accept "correct" outputs, instead of "incorrect" ones). However, appropriate reliance is not just a measure of how often users accept correct outputs but also how often they reject incorrect ones. This remains particularly challenging to measure in real-world applications of AI, especially GenAI systems.

A more practical approach to evaluating appropriate reliance is to assess the performance of a human+AI team compared to either a human or AI working alone. In this approach, appropriate reliance happens when the human+AI team performs better than the human or AI working alone. While less nuanced than AoR, a performance-driven approach to appropriate reliance provides an easier way to assess appropriate reliance. However, as the following subsection explains, such an approach raises a pertinent challenge.

3.2 A Strategy-Graded Approach to Appropriate Reliance

A big challenge for appropriate reliance efforts is that it is not always obvious when users over-rely on AI. Imagine an AI system that does tasks better than humans (AI accuracy > human accuracy). Even if users fully rely on this system (regardless of the correctness of its outputs), they will make more accurate decisions than working alone. The human+ AI team performs better than the human working alone. At first glance, this situation may seem acceptable and unproblematic. However, all AI systems make mistakes. Sometimes, AI systems make mistakes that are different from those humans make. Sometimes, AI systems make new types of

mistakes after receiving model updates (Bansal et al., 2019b). Thus, even when humans perform well using AI, the unpredictability of AI mistakes warrants caution against overreliance and a simple accuracy-focused approach to evaluating it.

Fok and Weld (2023) address this issue by distinguishing between outcome-graded and strategy-graded approaches to appropriate reliance. An *outcome-graded approach* to appropriate reliance focuses on the (in)correctness of the interaction outcome between users and AI (e.g., did the user accept the right AI outputs and reject the wrong ones?). Most research studies take an outcome-graded approach to appropriate reliance. However, as described above, the outcome-graded approach is susceptible to the challenge of identifying overreliance in situations where AI performs better than humans. It also fails to capture GenAI's complexity (e.g., that they can generate partially (in)correct outputs).

One possible solution, according to Fok and Weld (2023), is to take a *strategy-graded approach* to appropriate reliance. Under a strategy-graded approach, appropriate reliance happens when users accept AI outputs when the AI is expected to outperform users in a task and reject AI outputs when the AI is expected to underperform users in a task:

> Consider a decision-making task in which the human is historically 60% accurate, while the AI is 99.999% accurate. On any given instance of the task, if the human is uncertain of the answer, is it appropriate to rely on the AI's recommendation? Intuitively, the answer seems a clear 'yes'. But if the AI is later found incorrect, the outcome-graded definition says 'Inappropriate,' while the strategy-graded definition matches intuition and says 'Appropriate.' (Ibid., p. 8)

The strategy-graded approach makes clear the vital importance of realistic mental models for appropriate reliance. A mental model represents a user's understanding of the different aspects of a system, including how it works. Facilitating correct mental models of AI is central to enabling and facilitating appropriate reliance (Liao & Wortman Vaughan, 2024). For example, a person with a realistic mental model of a car can likely ascertain when it is seemingly working in order and acting up, even when they do not fully understand how a car works.

Fostering a strategy-graded approach to appropriate reliance, however, is challenging. Users, and at times even developers, struggle to develop effective mental models of *how* AI systems work, let alone *when* they are likely to be (in)correct. This is especially true in the context of GenAI. Identifying effective ways to operationalize a strategy-graded approach to appropriate reliance is thus a key research priority that involves working on multiple fronts, such as user education, UI messaging, UX interactions, and benchmarking and communicating domain- and task-specific AI performance.

4 Case Study: Overreliance on GenAI in Legal Practice

This section presents a case study to illustrate the issue of overreliance on AI in legal practice, a high-stakes domain where accuracy is crucial. The case study has three parts: (1) an example of an incident of overreliance on AI in legal practice, (2) a brief

analysis of potential reasons for why lawyers over-rely on AI, and (3) the use of uncertainty expressions as a possible mitigation for overreliance in the legal practice of case law research.

4.1 The Incident: A Lawyer Filed a Legal Affidavit that Cited Two Fake Court Cases

In February 2024, a Canadian lawyer submitted a set of legal briefs to the court for an ongoing case about a family dispute involving a parent's application to bring their children on a trip to China (Little, 2024). The problem was that one of the submitted legal affidavits cited two fictitious court cases. Here is a summary of one of the fictitious court cases (as described in the court's ruling):

> In M.M. v. A.M., 2019 BCSC 2060, the court granted the mother's application to travel with the child, aged 7, to India for six weeks to visit her parents and extended family. [. . .] The court found that the trip was in the best interests of the child, as it would allow him to maintain his cultural and familial ties, and that the mother had taken reasonable steps to address the father's concerns about the child's safety and health. The court also noted that the mother had provided a detailed travel itinerary, a consent letter from the father, and a return ticket for the child. (Ibid.)

Citing incorrect information in legal briefs is a serious issue because it indicates negligence, at best, and an intent to deceive the court, at worst. As the judge later ruled, it was the former. The two cases were generated by ChatGPT—a GenAI tool that the lawyer used in drafting the affidavit. The lawyer acknowledged that she had been "naïve about the risks of using ChatGPT" (Ibid.) and had not verified the cases provided by it. The court reprimanded the lawyer, ordered her to pay court fees for the additional effort and time, and overall criticized the uninformed use of AI tools in legal practice:

> As this case has unfortunately made clear, generative AI is still no substitute for the professional expertise that the justice system requires of lawyers. Competence in the selection and use of any technology tools, including those powered by AI, is critical. [. . .] The integrity of the justice system requires no less. (Ibid.)

This case is a textbook example of overreliance on AI, but it is by no means an exception. In the past 2 years alone, several cases of overreliance on GenAI in legal practice have surfaced in the states of New York (Merken, 2023), Texas (Merken, 2024), and Wyoming (Merken, 2025) and even internationally in countries such as Australia (Taylor, 2025) and India (Sharma, 2023).

4.2 The Problem: Why Do Lawyers Over-Rely on GenAI?

At first glance, it may seem confusing why trained lawyers cannot catch obvious AI mistakes. It seems easy to blame the lawyers. If lawyers had known better,

overreliance would not have happened. However, overreliance on AI is not just a human issue but a human-AI interaction and collaboration issue. As explained in Sect. 2, overreliance results from a complex interplay of sociotechnical factors—antecedents and mechanisms—making it difficult for both amateurs and experts to develop appropriate reliance on AI. Indeed, a close analysis of different instances of overreliance in legal practice reveals at least two common threads:

(a) **Lawyers Lack a Realistic Mental Model of GenAI**
Several lawyers who over-relied on GenAI tools acknowledged that they did not fully understand how such tools worked (Merken, 2025; Taylor, 2025). Some did not know GenAI tools can fabricate information (Moran, 2023), while others saw GenAI tools as "super-charged search engine[s]" (Weiss, 2024). Such lawyers were also unaware of the fact that a GenAI model's verbalized confidence does not accurately reflect the correctness of its output. Steven A. Schwartz, a New York lawyer, said in his defense that ChatGPT not only provided factitious legal sources but also "assured him of the reliability of the opinions and citations" that it provided (Ibid.).

(b) **The Form and Content of GenAI Outputs Inspire Unwarranted Trust in Them**
As described in Sect. 1.3, GenAI tools can generate volumes of impressive, novel content fast and effortlessly. GenAI outputs also often contain structural details associated with reliable source material, such as clear and descriptive headings, detailed lists, and well-formatted tables. Taken together, the form and content of GenAI outputs can make plausible but fake details seem true (Merken, 2025). The two fake GenAI-generated court cases in the Canadian case looked "100% real" to lawyers with detailed case facts, even though no such cases had ever been brought to court (Little, 2024).

Such factors, coupled with the reality that lawyers work under heavy workloads and time pressures, can easily lead to overreliance on AI. Confirmation bias only makes matters worse. Lawyers constantly try to find precedents to support their arguments, but such precedents may not always exist. However, when GenAI tools such as ChatGPT "magically" provide support for legal argumentation, it can lead lawyers to mistakenly assume they have finally found the needles in the haystack.

4.3 A Possible Solution: Uncertainty Expressions

Effectively addressing overreliance on AI in a legal practice (e.g., case law research) requires efforts on several fronts, including the organizational (e.g., establishing stricter verification processes), professional (e.g., standardizing the use of GenAI in legal research), algorithmic (e.g., improving the accuracy of GenAI tools), and design level (e.g., helping users form realistic mental models). In this subsection, we explore the use of a specific design-level mitigation for addressing overreliance in case law research.

Before we proceed, consider: what if we applied the *Overreliance on AI: Risk Identification and Mitigation Framework* (Microsoft, 2025) to this problem? A full demonstration of the framework's application is outside the chapter's scope, but a brief look at its application helps highlight its utility. Regarding overreliance risk identification, the framework can surface different *AI mistakes* (e.g., missing cases vs. fabricated sources), *negative consequences* of overreliance (e.g., productivity loss vs. professional cost), and its disparate impacts across *users* (e.g., early-career vs. experienced lawyers) and *use-cases* (e.g., case preparation vs. exploratory research). Regarding overreliance mitigation, the framework can outline different ways to *create realistic mental models* (e.g., put disclaimers about likely AI mistakes vs. explain AI limitations in user onboarding), *make it easy to spot AI mistakes* (e.g., show confidence scores next to AI output vs. use uncertainty expressions in AI outputs), and *facilitate verification* (e.g., make sources in AI outputs clickable vs. provide explanations for AI outputs). Practitioners can prioritize between different overreliance risk factors, selecting and assessing appropriate mitigations, aligned with identified risk and product goals. Overall, the framework translates the abstract issue of overreliance into tractable problems, intentional questions, and actionable decisions.

In the rest of this subsection, we will explore the use of *visual uncertainty expressions,* covered in subsection "Expressing Uncertainty by Highlighting Tokens in GenAI Outputs," as a specific design-level solution for addressing overreliance in case law research.

Consider Google's GenAI offering, Gemini. While there is research on highlights that convey token generation and edit probabilities (Spatharioti et al., 2023; Vasconcelos et al., 2023), Gemini uses a different kind of highlight. Users can use Gemini's "double-check response" feature to verify Gemini's output. The feature conducts Google Searches for different statements in Gemini's output. If the search results in content similar to a statement, Gemini highlights it in green. If it finds content different from a statement or does not find relevant content, it highlights it in red. Gemini leaves a statement unhighlighted if there isn't enough information to evaluate it.

Figure 5 is a screenshot of Gemini's color-coded output for a simulated prompt related to legal practice: *"Give me a list of court cases that help support the argument that an organization can fire an employee at will."* The output contains several unhighlighted and green-highlighted statements. It also contains a red-highlighted statement. A red highlight does not necessarily imply that the court case mentioned in the statement is fictitious; it does help convey to the user that a Google Search either found contradictory evidence for or could not triangulate the information in the statement.

However, like most mitigations of overreliance, such highlights can backfire. Consider an output with no red highlights, some green highlights, and several unhighlighted statements. The absence of red highlights may make users perceive that nothing is wrong with the output. This is not necessarily true. For example, an unhighlighted statement may contain inaccuracies that Gemini could not find sufficient evidence for with a Google Search. Further research is thus needed to

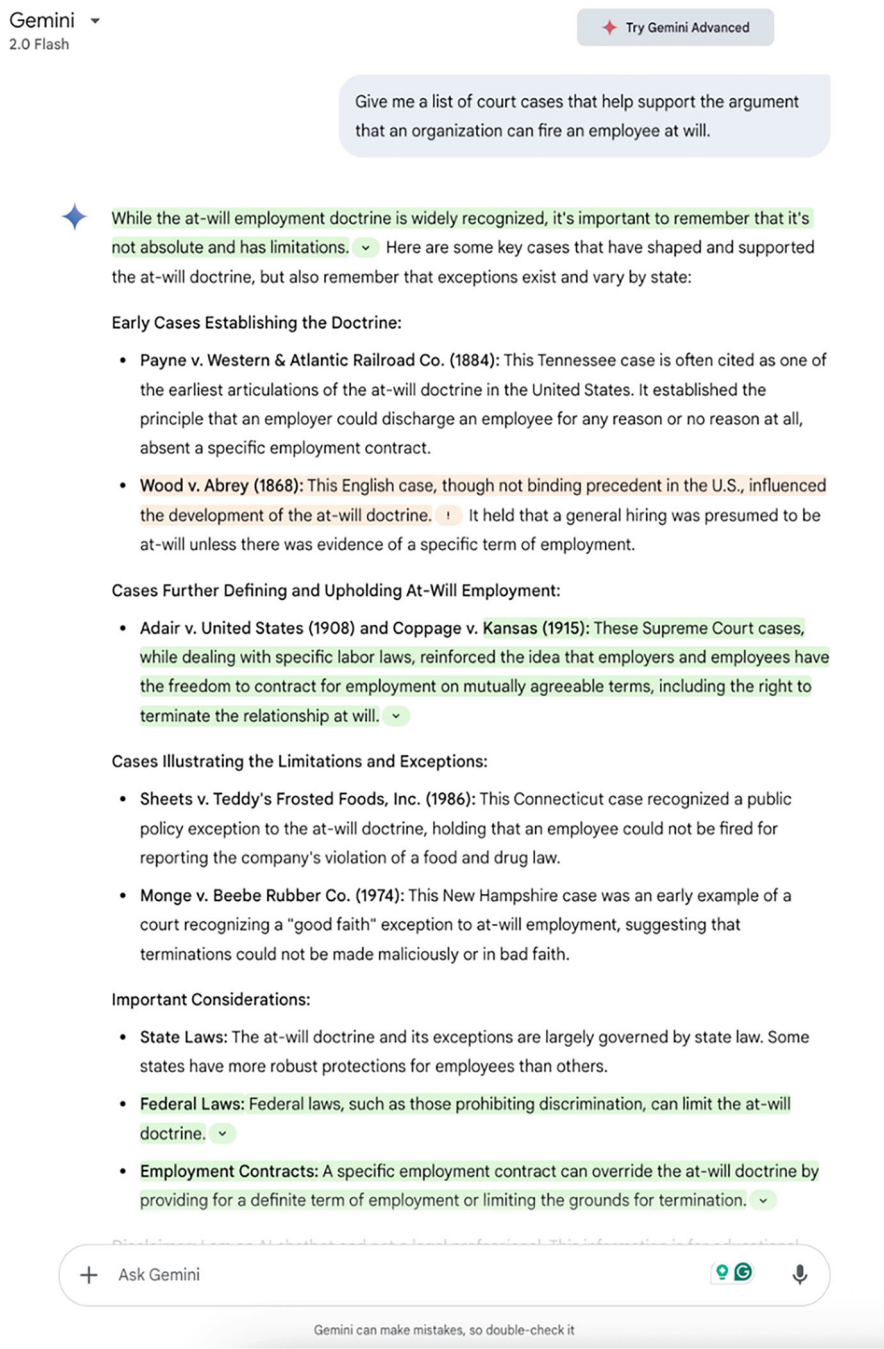

Fig. 5 Uncertainty expressions in Google Gemini outputs

investigate the use of such uncertainty highlights for different GenAI use cases in various professional practices.

5 Future Trends and Research Directions

Emerging research has begun to shed light on the issue of overreliance on GenAI. However, significant challenges remain, highlighting meaningful avenues for future research such as the following.

5.1 Systematically Investigating the Efficacy of Overreliance Mitigations

Future research should investigate the efficacy of different mitigations for overreliance across user groups, use cases, and application domains. Existing research shows that overreliance mitigations can sometimes backfire. For example, as explained in Sect. 2.4, interventions such as AI explanations may increase trust in AI outputs, even when the explanations are uninformative or wrong. Additionally, as explained in Sects. 2.1 and 2.2, the nature of overreliance on AI depends on individual differences and task characteristics. Contextual factors thus influence how and why users over-rely on AI, likely affecting the efficacy of overreliance mitigations across user groups, use cases, and application domains.

5.2 Exploring How the Modality of Human-AI Interaction Influences Overreliance on AI

Future research should explore the nature of overreliance on AI in different modalities of human-AI interactions, such as text-based, voice-based, and multimodal interactions. Existing research primarily focuses on text-based human-AI interaction. Different modalities, however, can shape user perceptions of AI in specific ways (e.g., voice-based interactions may further promote anthropomorphism) and raise new challenges for appropriate reliance on AI (e.g., how to present grounding information in voice-based interactions). Multimodal systems that blend text, audio, and video interactions further complicate matters (e.g., ChatGPT's advanced voice mode with video capability). As novel GenAI systems, including multimodal ones, find their way into users' personal lives and professional work, the impact of different modalities and their combinations on user trust in AI is largely unknown.

5.3 Investigating Overreliance on AI Agents and Multi-Agents

Future research should investigate the dynamics of overreliance in the context of autonomous AI agents. Existing research primarily focuses on AI systems that

respond to users and not systems that *act on behalf* of users. AI agents decide what to do next on their own without step-by-step user instructions; they make plans, use external tools, and act on the user's behalf. For example, instruct a travel-planning AI agent to "book a 1-week vacation in Finland with a budget of $2500," and it will search for flights and hotels, compare prices, buy the flight tickets and make hotel bookings, and even make restaurant reservations and purchase tour tickets. In short, AI agents are *expected* to work on their own *without* human oversight. This creates troubling conditions for overreliance. Users delegate entire workflows to agents, facing new challenges in forming effective mental models and in identifying the existence and provenance of errors and mistakes, especially in multi-agent systems where several AI agents interact and work together. Future research should examine how users work with and oversee AI agents, and make efforts to develop new interaction paradigms to support human oversight in human-agent interaction.

6 Conclusion

Overreliance on AI is a key issue for human-centered AI design, development, and deployment, stressing the urgent need to design AI tools with appropriate reliance in mind. The chapter began by highlighting the harmful effects of overreliance on AI, including poor human+AI team performance and ineffective human oversight. It described how antecedents of overreliance, such as low AI literacy and high expertise, and mechanisms, such as automation bias and overestimation of explanations, impact why people over-rely on AI and to what extent. The chapter outlined prominent mitigations for overreliance, such as explanations and cognitive forcing functions. While overreliance mitigations can often backfire, well-crafted mitigations such as self-critiques and uncertainty highlights show promise. A human-centered approach to addressing overreliance on AI emphasizes the importance of appropriate reliance to help users leverage AI's strengths and compensate for its weaknesses. The case study on AI mistakes in a high-stakes domain, such as legal practice, illustrated the real-world implications of overreliance, the place for human expertise, and the affordances and limitations of existing overreliance mitigations.

Competing Interest Declaration The author(s) has no competing interests to declare that are relevant to the content of this manuscript.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Shedding metacognitive light on reasoning research. *Trends in Cognitive Sciences, 21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)* (pp. 1–13). Association for Computing Machinery, Paper 3. https://doi.org/10.1145/3290605.3300233

Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., Chami, I., Sala, F., & Ré, C. (2022). Ask Me Anything: A simple strategy for prompting language models. arXiv preprint, arXiv: 2210.02441. https://doi.org/10.48550/arXiv.2210.02441

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H. S., Fernández, R., Plank, B., Sennrich, R., Zerva, C., & Aziz, W. (2023). Uncertainty in natural language generation: From theory to applications. arXiv preprint, arXiv:2307.15703. https://doi.org/10.48550/arXiv.2307.15703

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019a). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 19.

Bansal, G., Nushi, B., Kamar, E., Weld, D., Lasecki, W., & Horvitz, E. (2019b). A case for backward compatibility for human-AI teams. ArXiv:1906.01148 [Cs, Stat]. http://arxiv.org/abs/1906.01148

Bansal, G., Tongshuang, W. U., Zhou, J., Raymond, F. O. K., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). 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 sys*tems (pp. 1–16). Association for Computing Machinery, Article 81. https://doi.org/10.1145/3411764.3445717

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. (2021). Is the most accurate AI the best teammate? Optimizing AI for teamwork. *AAAI 2021*. https://www.microsoft.com/en-us/research/publication/is-the-most-accurate-ai-the-best-teammate-optimizing-ai-for-teamwork/

Bhatt, U., Antorán, A., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A., (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21*). Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

Biden, J. (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. *United States White House, Presidential Actions*, October 30, 2023.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). Measuring progress on scalable oversight for large language models. arXiv preprint, arXiv:2211.03540. https://doi.org/10.48550/arXiv.2211.03540

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 454–464). https://doi.org/10.1145/3377325.3377498

Buçinca, Z., Malaya, M. B., & 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), 188:1–188:21. https://doi.org/10.1145/3449287

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 international conference on healthcare informatics* (pp. 160–169). https://doi.org/10.1109/ICHI.2015.26

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–24. https://doi.org/10.1145/3359206

Chen, Z., & Chan, J. (2023). Large language model in creative work: The role of collaboration modality and user expertise. Available at SSRN: https://doi.org/10.2139/ssrn.4575598

Chen, V., Liao, Q. V., Wortman Vaughan, J., & 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), 370, 32 pages. https://doi.org/10.1145/3610219

Chiang, C.-W., & Yin, M. (2021). You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM web science conference 2021* (pp. 120–129). https://doi.org/10.1145/3447535.3462487

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research, 25*, e47184. https://doi.org/10.2196/47184

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023). Don't just tell me, ask me: AI systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal AI explanations. In *Proceedings of the 2023 CHI conference on human factors in computing systems (CHI '23)*. 13 p. https://doi.org/10.1145/3544548.3580672

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–12). https://doi.org/10.1145/3313831.3376638

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality* (Harvard Business School Technology & Operations Mgt. Unit Working Paper, No. 24-013). https://doi.org/10.2139/ssrn.4573321

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 275–285). https://doi.org/10.1145/3301275.3302310

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024a). The who in XAI: How AI background shapes perceptions of AI explanations. In *Proceedings of the 2024 CHI conference on human factors in computing systems (CHI '24)* (pp. 1–32). Association for Computing Machinery, Article 316. https://doi.org/10.1145/3613904.3642474

Ehsan, U., Liao, Q. V., Passi, S., Riedl, M. O., & Daumé, H. (2024b). Seamful XAI: Operationalizing seamful design in explainable AI. *Proc. ACM Hum.-Comput. Interact. 8, CSCW1,* Article 119 (April 2024), 29 pages. https://doi.org/10.1145/3637396

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society, 5*, 40–60. https://doi.org/10.17351/ests2019.260

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, *Pub. L. No. COM(2021) 206 final*. https://op.europa.eu/s/z5Td

Fok, R., & Weld, D. S. (2023). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. arXiv preprint, arXiv:2305.07722. https://doi.org/10.48550/arXiv.2305.07722

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *Npj Digital Medicine, 4*(1), 1–8. https://doi.org/10.1038/s41746-021-00385-9

Goh, E., Gallo, R., Hom, J., et al. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open, 7*(10). https://doi.org/10.1001/jamanetworkopen.2024.40969

Goyal, N., Briakou, E., Liu, A., Baumler, C., Bonial, C., Micher, J., Voss, C. R., Carpuat, M., & Daumé, H. (2023). What else do I need to know? The effect of background information on users' reliance on QA systems. arXiv preprint, arXiv:2305.14331. https://doi.org/10.48550/arXiv.2305.14331

Green, B. (2021). *The flaws of policies requiring human oversight of government algorithms* (SSRN Scholarly Paper ID 3921216). Social Science Research Network. https://doi.org/10.2139/ssrn.3921216

Green, B., & Chen, Y. (2019a). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 50:1–50:24. https://doi.org/10.1145/3359152

Green, B., & Chen, Y. (2019b). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 90–99). https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. ArXiv:2012.05370 [Cs]. https://doi.org/10.1145/3479562

Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency (FAccT '24)* (pp. 221–236). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658901

Jackson, S., Liu, J., Singh, R., & Passi, S. (2025). Maintaining data infrastructures. In T. Venturini, A. Acker, J. Plantin, & T. Walford (Eds.), *The Sage handbook of data and society*. Sage. https://sk.sagepub.com/hnbk/edvol/the-sage-handbook-of-data-and-society/toc

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational Psychiatry, 11*(1), 1–9. https://doi.org/10.1038/s41398-021-01224-x

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12). https://doi.org/10.1145/3571730

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3313831.3376219

Kim, A., Yang, M., & Zhang, J. (2020). *When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms* (SSRN Scholarly Paper ID 3691575). Social Science Research Network. https://doi.org/10.2139/ssrn.3691575

Kim, S., Liao, V., Vorvoreanu, M., Ballard, S., & 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 (FAccT '24)* (pp. 822–835). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658941

Koulu, R. (2020). Human control over automation: EU policy and AI ethics. *European Journal of Legal Studies, 12*(1), 9–46.

Krishna, S., Agarwal, C., & Lakkaraju, H. (2024). Understanding the effects of iterative prompting on truthfulness. arXiv preprint, arXiv: 2402.06625. https://doi.org/10.48550/arXiv.2402.06625

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38). https://doi.org/10.1145/3287560.3287590

Lai, V., Liu, H., & Tan, C. (2020). "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376873

Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety, 25*(10), 808–820.

Lee, M. K., & Rich, K. (2021). Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445570

Levy, A., Agrawal, M., Satyanarayan, A., & Sontag, D. (2021). Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445522

Li, Z., Lu, Z., & Yin, M. (2023). Modeling human trust and reliance in AI-assisted decision making: A Markovian approach. In *Proceedings of the thirty-seventh aaai conference on artificial intelligence and thirty-fifth conference on innovative applications of artificial intelligence and thirteenth symposium on educational advances in artificial intelligence (AAAI'23/IAAI'23/EAAI'23)* (vol. 37, pp. 6056–6064). AAAI Press, Article 679. https://doi.org/10.1609/aaai.v37i5.25748

Liang, J. T., Yang, C., & Myers, B. A. (2024). A large-scale survey on the usability of AI programming assistants: Successes and challenges. In *Proceedings of the 46th IEEE/ACM international conference on software engineering* (p. 13). https://doi.org/10.1145/3597503.3608128

Liao, Q. V., & Wortman Vaughan, J. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8036d03b

Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. arXiv preprint, arXiv:2205.14334. https://doi.org/10.48550/arXiv.2205.14334

Little, S. (2024). AI 'hallucinated' fake legal cases allegedly filed to B.C. court in Canadian first. *Global News.* https://globalnews.ca/news/10238699/fake-legal-case-bc-ai/

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Lu, Z., & 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* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445562

Merken, S. (2023). New York legal scene faces tests after tumultuous 2023. Reuters. https://www.reuters.com/legal/transactional/new-york-legal-scene-faces-tests-after-tumultuous-2023-2023-12-28/

Merken, S. (2024). Texas lawyer fined for AI use in latest sanction over fake citations. Reuters. https://www.reuters.com/legal/government/texas-lawyer-fined-ai-use-latest-sanction-over-fake-citations-2024-11-26/

Merken, S. (2025). Lawyers in Walmart lawsuit admit AI 'hallucinated' case citations. Reuters. https://www.reuters.com/legal/legalindustry/lawyers-walmart-lawsuit-admit-ai-hallucinated-case-citations-2025-02-10/

Microsoft. (2021). Microsoft HAX Toolkit: Hands-on tools for building effective human-AI experiences. *Microsoft.* https://www.microsoft.com/en-us/haxtoolkit/

Microsoft. (2025). Overreliance on AI: Risk identification and mitigation framework. *Microsoft.* https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/overreliance-on-ai/overreliance-on-ai

Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y.-L. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics, 10*, 857–872. https://doi.org/10.1162/tacl_a_00494

Moran, L. (2023). Lawyer cites fake cases generated by ChatGPT in legal brief. Legal Dive. https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. ArXiv:1802.00682 [Cs]. http://arxiv.org/abs/1802.00682

Nourani, M., King, J., & Ragan, E. (2020). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8*, 112–121.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *26th international conference on intelligent user interfaces* (pp. 340–350). https://doi.org/10.1145/3397481.3450639

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. ArXiv:1907.12652 [Cs]. http://arxiv.org/abs/1907.12652

Passi, S., & Jackson, S. J., (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proc. ACM Hum.-Comput. Interact*. 2, CSCW, 136, 28. https://doi.org/10.1145/3274405

Perez, E., & Long, R. (2023). Towards evaluating AI systems for moral status using self-reports. arXiv preprint, arXiv:2311.08576. https://doi.org/10.48550/arXiv.2311.08576

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., . . . Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. arXiv preprint, arXiv:2212.09251. https://doi.org/10.48550/arXiv.2212.09251

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors, 57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–52). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445315

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2023). "It's Weird that it knows what I want": Usability and interactions with copilot for novice programmers. *ACM Transactions on Computer-Human Interaction, 31*(*1*). https://doi.org/10.1145/3617367

Radensky, M., Séguin, J. A., Lim, J. S., Olson, K., & Geiger, R. (2023). "I think you might like this": Exploring effects of confidence signal patterns on trust in and reliance on conversational recommender systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 792–804). https://doi.org/10.1145/3593013.3594043

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., . . . Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. arXiv preprint, arXiv: 2110.08207. https://doi.org/10.48550/arXiv.2110.08207

Sarkar, A. (2024). AI should challenge, not obey. *Communications of the ACM (CACM)*. Forthcoming. https://www.microsoft.com/en-us/research/uploads/prod/2024/03/CACM_2024_AI_provocateur_MSFT_internal_preprint.pdf

Sarkar, A., Gordon, A. D., Negreanu, C., Poelitz, C., Ragavan, S. S., & Zorn, B. (2022). What is it like to program with artificial intelligence? In *Proceedings of the 33rd annual conference of the Psychology of Programming Interest Group (PPIG 2022).* https://doi.org/10.48550/arXiv.2208.06213

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. arXiv preprint, arXiv: 2206.05802. https://doi.org/10.48550/arXiv.2206.05802

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 240–251). https://doi.org/10.1145/3301275.3302308

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & 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* (pp. 410–422). https://doi.org/10.1145/3581641.3584066

Sellen, A., & Horvitz, E. (2023). The rise of the AI Co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM, 67*, 7 (July 2024), 18–23. https://doi.org/10.1145/3637865

Sharma, D. (2023). Lawyer faces trouble after using ChatGPT for research, AI tool comes up with fake cases that never existed. India Today. https://www.indiatoday.in/technology/news/story/lawyer-faces-trouble-after-using-chatgpt-for-research-ai-tool-comes-up-with-fake-cases-that-never-existed-2385542-2023-05-28

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint, arXiv:2310.13548. https://doi.org/10.48550/arXiv.2310.13548

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daumé III, H., & Boyd-Graber, J. (2023). Large language models help humans verify truthfulness – Except when they are convincingly wrong. arXiv preprint, arXiv:2310.12558. https://doi.org/10.48550/arXiv.2310.12558

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment. arXiv preprint, arXiv:2307.03744. https://doi.org/10.48550/arXiv.2307.03744

Springer, A., Hollis, V., & Whittaker, S. (2017). Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI spring symposium series*. https://www.aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L., & Smyth, P. (2024). The calibration gap between model and human confidence in large language models. arXiv preprint, arXiv: 2401.13835. https://doi.org/10.48550/arXiv.2401.13835

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). Investigating explainability of generative AI for code through scenario-based design. In *27th international conference on intelligent user interfaces* (pp. 212–228). https://doi.org/10.1145/3490099.3511119

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. In *12th ACM conference on web science* (pp. 315–324). https://doi.org/10.1145/3394231.3397922

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). Investigating human + machine complementarity for recidivism predictions. arXiv:1808.09123 [Cs, Stat]. http://arxiv.org/abs/1808.09123

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. *In Proceedings of the 2024 CHI conference on human factors in computing systems (CHI'24)* (article 680, pp. 1–24). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642902

Taylor, J. (2025). Lawyer caught using AI-generated false citations in court case penalised in Australian first. The Guardian. https://www.theguardian.com/law/2025/sep/03/lawyer-caught-using-ai-generated-false-citations-in-court-case-penalised-in-australian-first

Topolinski, S., & Reber, R. (2010). Immediate truth – Temporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. *Cognition, 114*(1), 117–122. https://doi.org/10.1016/j.cognition.2009.09.009

Vaccaro, M., & Waldo, J. (2019). *The effects of mixing machine learning and human judgment*. https://cacm.acm.org/magazines/2019/11/240386-the-effects-of-mixing-machine-learning-and-human-judgment/fulltext

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., & Wortman Vaughan, J. (2023). Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. arXiv preprint, arXiv:2302.07248. https://doi.org/10.48550/arXiv.2302.07248

Vodrahalli, K., Daneshjou, R., Gerstenberg, T., & Zou, J. (2022). Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society (AIES '22)* (pp. 763–777). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534150

Vorvoreanu, M., Passi, S., Dhanorkar, S., Heger, A., & Walker, K. (2025). Fostering appropriate reliance on GenAI: Lessons learned from early research. *Microsoft Technical Report. MSR-TR-2025-4*. https://www.microsoft.com/en-us/research/publication/fostering-appropriate-reliance-on-genai-lessons-learned-from-early-research/

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. ArXiv.

Wason, P. C., & Evans, J. (1974). Dual processes in reasoning? *Cognition, 3*(2), 141–154.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., . . . Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 214–229). https://doi.org/10.1145/3531146.3533088

Weiss, D. C. (2024). Fake citations in legal brief were generated by Google Bard AI program, says ex-Trump lawyer Michael Cohen. *ABA Journal*. https://www.abajournal.com/news/article/fake-citations-in-legal-brief-were-generated-by-google-bard-ai-program-says-ex-trump-lawyer-michael-cohen

Weisz, J. D., Muller, M., Houde, S., Richards, J., Ross, S. I., Martinez, F., Agarwal, M., & Talamadupula, K. (2021). Perfection not required? Human-AI partnerships in code translation. In *26th international conference on intelligent user interfaces (IUI '21)* (pp. 402–412). Association for Computing Machinery. https://doi.org/10.1145/3397481.3450656

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design principles for generative AI applications. arXiv preprint, arXiv:2401.14484. https://doi.org/10.48550/arXiv.2401.14484

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv preprint, arXiv:2306.13063. https://doi.org/10.48550/arXiv.2306.13063

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions, 26*(4), 42–46. https://doi.org/10.1145/3328485

Xu, W., & Gao, Z. (2024). An intelligent sociotechnical systems (iSTS) framework: Enabling a hierarchical human-centered AI (hHCAI) approach. *IEEE Transactions on Technology and Society, 6*(1), 31–462025. https://doi.org/10.1109/TTS.2024.3486254

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *International Journal of Human–Computer Interaction, 39*(3), 494–518. https://doi.org/10.1080/10447318.2022.2041900

Xu, W., Gao, Z., & Dainoff, M. (2024). An HCAI methodological framework (HCAI-MF): Putting it into action to enable human-centered AI. arXiv preprint, arXiv: 2311.16027. https://doi.org/10.48550/arXiv.2311.16027

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making, 32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300509

Zhang, B., & Dafoe, A. (2019). *Artificial intelligence: American attitudes and trends*. Tech. rep., Centre for the Governance of AI, University of Oxford. 2 General attitudes toward AI|Artificial Intelligence: American Attitudes and Trends (governanceai.github.io). https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/general-attitudes-toward-ai.html

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (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* (pp. 295–305). https://doi.org/10.1145/3351095.3372852

Zhou, K., Jurafsky, D., & Hashimoto, T. (2023). Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 5506–5524). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.335

Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the unreliable: The impact of language models' reluctance to express uncertainty. arXiv preprint, arXiv:2401.06730. https://doi.org/10.48550/arXiv.2401.06730