

Epistemic Trustworthiness in Generative AI: A Normative Framework for Warranted Reliance in High-Stakes Workflows

Nimisha Karnatak, Max Van Kleek, Nigel Shadbolt

Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom
nimisha.karnatak@some.ox.ac.uk, max.vankleek@cs.ox.ac.uk, nigel.shadbolt@jesus.ox.ac.uk

Abstract

Generative AI systems are increasingly deployed in high-stakes professional contexts, where their outputs shape what users believe, how they reason, and what they treat as settled. This raises a central question for responsible AI: under what conditions is reliance on generative AI outputs epistemically warranted rather than behaviourally induced? Existing frameworks largely ask whether AI outputs are accurate, fair, explainable, safe, or trusted by users. These questions remain necessary, and each can contribute to warranted reliance. However, they do not directly specify warranted reliance as a distinct evaluative target: the conditions under which users are justified in treating AI outputs as inputs into their own reasoning. We argue that this requires an account of epistemic trustworthiness: what makes a system epistemically worthy of reliance. Drawing on philosophical accounts of trustworthiness as competence and audience-orientation, we develop a constitutive normative framework comprising three jointly necessary and non-fungible conditions. First, epistemic humility requires systems to represent and communicate the limits of their competence. Second, epistemic access requires systems to enable users to inspect, question, and contest outputs in context. Third, resistance to epistemic injustice requires systems to recognise users as legitimate epistemic agents and avoid marginalising their knowledge and experience. Through real-world case analyses in legal reasoning, medical reasoning, and hiring, we show how failures of epistemic humility, epistemic access, and resistance to epistemic injustice can produce consequential harms that standard measures of accuracy, fairness, and usability do not address on their own. We conclude by outlining design and evaluation implications for GenAI systems organised around epistemically warranted reliance rather than output correctness alone.

1 Introduction

Generative AI systems are increasingly deployed as epistemic infrastructure in institutional knowledge work, reshaping how knowledge is created, synthesised, and acted upon in professional settings (Marchal et al. 2026; Appel et al. 2025; Yun et al. 2025). Professionals in high-stakes settings such as clinical decision support, legal guidance, hiring, and policy advice (Sivaraman et al. 2023) now rely on these systems to form beliefs, make decisions, and coordinate action.

In such contexts, reliance on system outputs constitutes a form of epistemic deference: users treat generated claims, summaries, recommendations, or explanations as inputs into their own reasoning (Lange 2026; Slome 2026; Chen et al. 2023). Whether such deference is warranted, and therefore whether the system is epistemically trustworthy, is not a peripheral concern but a precondition of responsible AI deployment (Liao and Sundar 2022)¹.

Existing approaches to responsible AI have made important progress in evaluating system accuracy, fairness, safety, calibration, explainability, and appropriate reliance (Mehrotra et al. 2024). Yet these approaches do not fully specify when reliance on a particular GenAI output is epistemically warranted. A system may produce outputs that are accurate, fluent, source-grounded, or explainable while still failing to provide users with adequate grounds to rely on, verify, contest, or withhold reliance from those outputs (Mehrotra et al. 2024). We therefore ask: what must a generative AI system provide, at the moment of interaction, for users' reliance on its outputs to be epistemically warranted?

We answer this question by developing a constitutive normative framework of epistemic trustworthiness for generative AI (Koskinen 2024). Drawing on social epistemology and philosophical accounts of trust and testimony (Baier 1986; Scheman 2001; Lackey 2008), we define epistemic trustworthiness as a property of the user–system relation: a generative AI system is epistemically trustworthy when it provides the conditions under which a situated user has adequate grounds to rely on, verify, contest, or withhold reliance from its outputs. The framework specifies three jointly necessary and non-fungible conditions. First, *epistemic humility* requires systems to represent and communicate the limits of their competence (Sosa 2007). Second, *epistemic access* requires systems to enable users to inspect, question, and contest outputs in context (Sterz et al. 2024). Third, *resistance to epistemic injustice* requires systems to recognise users and affected communities as legitimate epistemic agents, rather than marginalising their knowledge, concepts, or experience (Kay, Kasirzadeh, and Mohamed 2024).

The framework treats epistemic trustworthiness as constitutive rather than merely regulative. Existing concepts such

as calibration, explainability, provenance, abstention, and fairness identify mechanisms that may support trustworthy systems (Sanderson, Douglas, and Lu 2023). Our framework instead specifies the non-fungible conditions that must jointly hold for reliance on generative AI outputs to be epistemically warranted.

We make two contributions. First, drawing on social epistemology and philosophical work on trust and testimony, we develop a constitutive framework for warranted reliance on generative AI by translating competence and audience-orientation into the GenAI setting (Baier 1986; Scheman 2001; Lackey 2008; DiMarco 2023). The framework specifies three jointly necessary and non-fungible conditions: epistemic humility, epistemic access, and resistance to epistemic injustice. We further map their intervention points across the model, data, and interface layers of the sociotechnical stack, providing a layer-explicit structure for diagnosing where warranted reliance breaks down. Second, we demonstrate the framework’s diagnostic value by positioning it against existing responsible AI approaches and applying it to real-world cases, showing how failures of epistemic humility, epistemic access, and resistance to epistemic injustice can produce consequential harms even when systems appear accurate, usable, source-grounded, or safe under standard evaluative lenses.

2 Related Work

2.1 Why GenAI-Mediated Reliance Requires an Epistemic Framework

Generative AI creates a distinctive problem of epistemic reliance. Unlike many earlier decision-support systems that produce bounded predictions, classifications, or recommendations, generative AI systems produce open-ended claims, summaries, explanations, and analyses that users may incorporate into their own reasoning. This makes their outputs appear testimony-like: they are expressed in fluent natural language, often framed with apparent confidence, and increasingly embedded within institutional workflows (Anderl et al. 2024; Metzger et al. 2024). Prior work in HCI and communication shows that users may treat fluent and confident presentation as a cue of credibility, even when those cues do not reflect whether the output is accurate or adequately supported (Anderl et al. 2024; Metzger et al. 2024). Conceptual critiques of language models similarly caution that fluent text generation should not be mistaken for grounded reference, communicative intent, or belief (Chiesurin et al. 2023). The resulting problem is not merely that generative AI systems can be wrong, but that they can be wrong in ways that still invite epistemic deference.

This deference becomes difficult to justify because generative AI failures are often uneven, opaque, and context-dependent. Systems may hallucinate unsupported claims, express unwarranted confidence, misattribute evidence, omit relevant uncertainty, or produce responses that are locally fluent but contextually misaligned (Simhi et al. 2025). Such failures are closely related to what has been described as the jagged frontier of AI capability: models may perform impressively on some tasks while failing unexpectedly on

adjacent ones (Dell’Acqua et al. 2026; Morris et al. 2026). In professional settings, this unevenness is especially consequential because users may not know when a system has moved from competence to failure. A generated answer may appear complete, procedurally legitimate, or institutionally authoritative, even when the system has not provided adequate grounds for reliance.

Existing work on epistemic trustworthiness and trust in AI approaches this normative question from distinct starting points. Ferrario develops a reliabilist account on which users’ credences in a system’s trustworthiness are justified when generated by a reliable assessment process that tends to align perceived trustworthiness with actual trustworthiness, while Simon examines how multiple forms of GenAI-related deception produce misplaced trust and distrust and undermine the conditions for justified trust (Ferrario 2024; Simon 2026). Jonas, Greussing, and Taddicken analyse how users perceive trustworthiness across the interface, underlying infrastructure, developers, and organisations (Jonas, Greussing, and Taddicken 2025). Simion and Kelp ground AI trustworthiness in a sufficiently strong disposition to fulfil function-based obligations (Simion and Kelp 2023). Tanchuk and Taylor develop a shared-responsibility account spanning individuals, tools, institutions, and their epistemic environment, while Marchal et al. connect agent-level properties with provenance, governance, and wider socio-epistemic infrastructures (Tanchuk and Taylor 2025; Marchal et al. 2026). These accounts address user–system interaction to varying degrees, but none takes as its primary unit of analysis the situated relation between a user and a particular GenAI output. Nor do they provide a unified set of constitutive conditions under which that user has adequate grounds to rely on, verify, contest, or withhold reliance from the output. We address this gap by specifying the relational conditions required for epistemically warranted reliance on a particular GenAI output.

2.2 The Behavioural and Output-Centric Evaluative Gap

Human–AI interaction research often operationalises appropriate reliance behaviourally: users should accept correct algorithmic advice and reject incorrect advice (Lee and See 2004; Yin, Wortman Vaughan, and Wallach 2019). This tradition has identified failures such as automation complacency, in which users defer to faulty recommendations, and algorithm aversion, in which users reject superior automated advice after observing minor errors (Dietvorst, Simmons, and Massey 2015; Parasuraman and Manzey 2010). More recent work shows that confidence displays, explanations, and descriptions of model capabilities can shape reliance behaviour (Cau and Spano 2025; Zhang, Liao, and Bellamy 2020). Such evaluations reveal whether reliance corresponds to system correctness, but not its epistemic basis. For example, a user may accept a correct output after examining relevant evidence and uncertainty, or because interface cues encourage deference independently of output quality (Liao and Sundar 2022; Chen et al. 2023).

Responsible AI evaluation examines complementary properties of systems and outputs, including accuracy, ro-

Evaluative lens	Focus	Primary question
Behavioural evaluation	User reliance behaviour	Does reliance track output correctness?
System- and output-centred evaluation	System and output properties	Are relevant performance and governance criteria satisfied?
Epistemic-relational evaluation	Situated user–system relation	Does the situated user–system relation support epistemically warranted reliance on a particular output?

Table 1: Three evaluative lenses on reliance in generative AI. Behavioural evaluation and evaluations of system and output properties provide evidence relevant to warranted reliance; epistemic-relational evaluation assesses its relational conditions.

bustness, fairness, safety, calibration, explainability, privacy, and accountability (Bommasani, Liang, and Lee 2023; Kowald et al. 2024). Sociotechnical approaches further show that these properties must be evaluated within their social and institutional contexts (Selbst et al. 2019). These approaches address important dimensions of responsible AI, but they do not generally make warranted reliance within a particular user–output interaction their primary evaluative target. This distinction is especially important in professional knowledge work, where GenAI systems produce open-ended claims, explanations, summaries, and recommendations that users incorporate into their reasoning (Yun et al. 2025). System- and output-level evaluations may show that an AI satisfies relevant performance or governance criteria without showing whether users can identify its limitations, inspect the grounds for its outputs, contest those outputs, or have their relevant knowledge taken into account.

To address this gap, our framework takes the relation between a particular user and a particular output in context as its unit of analysis and specifies what must hold in that relation for reliance to be epistemically warranted.

3 Framework for Epistemic Trustworthiness in GenAI

We next formulate a framework of three conditions for epistemic trustworthiness in generative AI. The purpose of the framework is to translate philosophical accounts of epistemic trustworthiness into a form that can inform responsible AI evaluation and design. We intend the framework to function as a conceptual tool for analysing when reliance on GenAI outputs is epistemically warranted, diagnosing cases in which such reliance becomes unwarranted, and guiding the design of systems that preserve users’ epistemic agency: their capacity to calibrate deference, inspect and contest outputs, and appropriately rely on AI-generated claims.

3.1 Assumptions of the Framework

Before presenting the framework, we make explicit three assumptions that delimit the scope of the framework.

Scope. First, the framework is scoped to GenAI systems deployed in contexts where human epistemic agency is at stake: contexts in which system outputs actively shape what users come to believe, how they reason, and what they treat as settled (Marchal et al. 2026; Hila 2026). This includes large language models, multimodal systems,

retrieval-augmented generation, and multi-agent pipelines used in high-stakes professional settings such as clinical decision support (Sivaraman et al. 2023), legal guidance (Magesh et al. 2025), and policy advice (Karnatak et al. 2026). The framework is less applicable to systems whose outputs function only as low-stakes data points or operational signals. However, classification, recommendation, and decision-support systems may fall within scope when their outputs are treated as evidence, advice, or authoritative input for human judgement.

Normative status. Second, the framework is normative rather than fully operationalised. It specifies what epistemic trustworthiness requires of GenAI systems at the level of principle, rather than providing a validated measurement instrument or scoring rubric (Whittlestone et al. 2019). Determining what counts as adequate uncertainty communication, sufficient contestability, or genuine recognition of epistemic standing will often require context-sensitive judgement and empirical validation (Sanderson et al. 2023). The framework identifies what must be evaluated; it does not resolve every operationalisation decision (Mittelstadt 2019).

Instrumental stance. Third, we adopt an instrumental stance towards epistemic properties in AI systems (Dennett 1987). We do not assume that AI systems possess genuine epistemic states such as belief, knowledge, or self-awareness. Instead, we evaluate systems in terms of whether their outputs function as if they exhibit relevant epistemic properties: whether they effectively signal uncertainty, support justified belief formation, or recognise the epistemic standing of the user. This allows us to translate philosophical concepts into operational criteria without making claims about whether AI systems genuinely know, believe, or understand.

3.2 Deriving the Framework

We derive our normative framework by translating social epistemological accounts of trustworthy testimony into the GenAI setting. This translation is appropriate because GenAI systems produce fluent, open-ended natural-language assertions that users may treat as testimony-like inputs into their reasoning, rather than as structured data points to be independently evaluated (Heersmink et al. 2024). We use this framework instrumentally: without attributing beliefs, intentions, or moral agency to large language models, we ask what structural conditions must hold

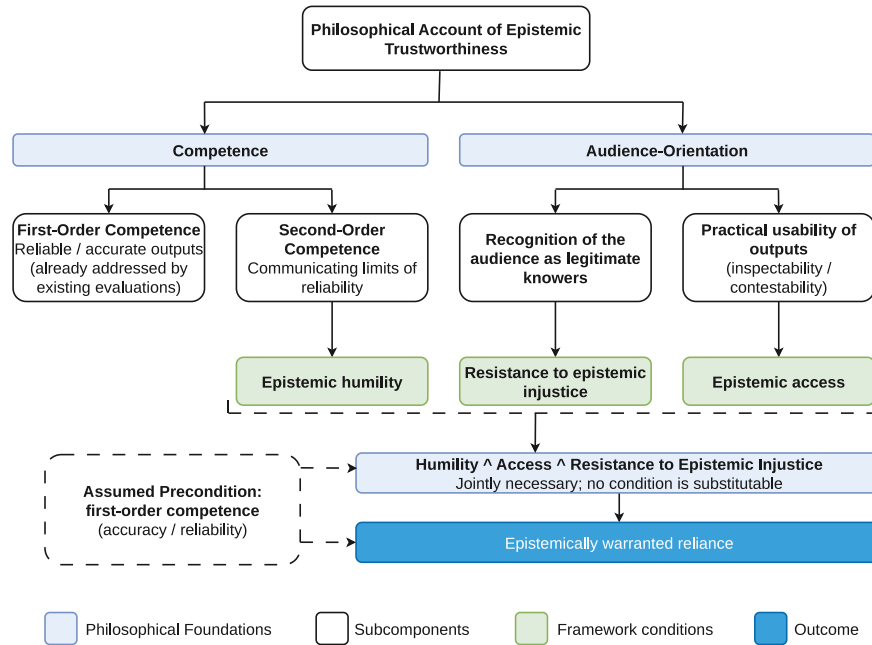


Figure 1: Derivation of the three conditions for epistemic trustworthiness in generative AI from philosophical accounts of competence and audience-orientation. The three framework conditions, namely epistemic humility, epistemic access, and resistance to epistemic injustice, are derived from second-order competence and the two dimensions of audience-orientation. Their joint satisfaction, presupposing first-order accuracy, constitutes the system-side condition for epistemically warranted reliance.

when their outputs function as epistemic inputs into professional reasoning. Within social epistemology, trustworthy testimony is commonly grounded in two components: *competence*, the capacity to produce well-supported claims, and *audience-orientation*, responsiveness to the epistemic situation, vulnerabilities, and needs of those who depend on the source (Baier 1986; Scheman 2001; Lackey 2008).

From Competence to Humility In AI evaluation, competence is often operationalised as first-order reliability: accuracy, calibration, or performance against ground-truth data (Bommasani, Liang, and Lee 2023; Raji et al. 2021). These measures are necessary but insufficient for epistemic trustworthiness. A source can produce reliable claims on average while still failing to recognise the limits of its own reliability in a specific case. Following accounts of second-order competence, epistemic trustworthiness therefore requires not only reliable output production but also the capacity to identify and communicate the conditions under which reliability is limited (Sosa 2007). We label the GenAI translation of this requirement *epistemic humility* and define it in Section 3.3.

From Audience-Orientation to Access and Recognition Audience-orientation requires that an epistemic source be responsive to the situated conditions of those who depend on it. In translating this requirement to GenAI, we distinguish two dimensions: a practical-communicative dimension and a recognitional dimension. The practical-communicative dimension concerns whether users can evaluate what the sys-

tem offers; the recognitional dimension concerns whether users and affected communities are treated as legitimate epistemic agents.

- **The Practical-Communicative Dimension (Epistemic Access).** For an output to support warranted reliance, it must be practically evaluable by the audience in context. This yields our second condition, *epistemic access*: the user’s practical ability to inspect, interpret, and contest the basis of a system output. This requirement draws on the account of the good informant developed by Craig (1990), in which trustworthy testimony requires not only competence but also a detectable indication that the inquirer can use to judge whether to rely. We formally define this condition in Section 3.3.
- **The Recognitional Dimension (Resistance to Epistemic Injustice).** Audience-orientation also presupposes that the audience is recognised as a legitimate epistemic agent. Feminist social epistemology shows that epistemic injustice occurs when people’s credibility, concepts, language, or experience are systematically downgraded or excluded (Fricker 2007; Medina 2013). Dotson (2011) further shows that testimonial exchange can fail when hearers fail to extend communicative reciprocity to speakers as legitimate knowers. The corresponding condition, *resistance to epistemic injustice*, is developed in Section 3.3.

3.3 Defining the Three Conditions

Epistemic humility. In social epistemology, competence involves more than producing true or well-supported claims; it also requires second-order competence: the capacity to recognise the limits of one's own reliability (Sosa 2007; Lackey 2008). Translated to GenAI, epistemic humility refers to a system's capacity to make those limits visible at the moment reliance is being formed (Griot et al. 2025). It is not equivalent to low confidence, generic disclaimers, or occasional abstention (Metzger et al. 2024; Wen, Howe, and Wang 2024). A system satisfies epistemic humility when it helps users understand whether an output is well grounded, uncertain, underspecified, outside the system's reliable scope, or in need of verification before use. Humility is therefore distinct from first-order accuracy: a system can be accurate yet fail humility (by overstating certainty about correct outputs) or inaccurate yet succeed at humility (by appropriately flagging fabricated outputs as unverified). Humility decomposes into two properties: actionable limitation-signalling (H1), whereby the system explains what the user should do in response to a material limitation rather than merely stating that uncertainty exists (Vasconcelos et al. 2023); and interactional humility (H2), whereby appropriate limitation-signalling is maintained across turns, including when users request confirmation, question the output, or provide new information (Sharma et al. 2024).

Epistemic access. Audience-orientation requires that those who depend on a source are able to evaluate its claims in practice (Craig 1990). Translated to GenAI, epistemic access refers to the user's practical ability to inspect, interpret, and contest the basis of a system output in context (Alfrink et al. 2023). It is not equivalent to the mere presence of explanations, citations, or documentation (Vasconcelos et al. 2023; Chen et al. 2023). A system satisfies epistemic access when users can determine what evidence supports a claim, how strongly that evidence supports it, what assumptions or omissions shape the output, and how the output can be questioned, revised, or overridden. Access decomposes into three properties: verifiable claim–evidence linkage (A1), inspectable retrieval (A2), and contestability (A3). Accordingly, providing citations does not by itself establish epistemic access. Users must be able to determine whether a source supports the specific claim, understand why it was selected, and challenge or correct the output when inspection reveals a problem.

Resistance to epistemic injustice. Audience-orientation also presupposes that the audience is recognised as a legitimate epistemic agent. In feminist social epistemology, epistemic injustice describes cases in which people's credibility, concepts, language, or experience are systematically downgraded or excluded (Fricker 2007; Dotson 2011; Medina 2013). Translated to GenAI, resistance to epistemic injustice refers to a system's capacity to avoid marginalising the knowledge of particular users or communities (Kay, Kasirzadeh, and Mohamed 2024). It is not equivalent to aggregate fairness alone (Selbst et al. 2019). A system satisfies this condition when users and affected communities are recognised as legitimate knowers whose testimony, concepts, language, and contextual knowledge can shape what

the system treats as credible, relevant, or actionable. Resistance decomposes into two properties: equal credibility weighting across identity-signalling features (R1) (Fricker 2007; Wilson and Caliskan 2024; Kay, Kasirzadeh, and Mohamed 2024), and recognitional adequacy (R2) (Kay, Kasirzadeh, and Mohamed 2024), under which users outside institutionally privileged roles are treated as legitimate knowers whose need for guidance is taken seriously.

The three conditions are conceptually parallel, but they have distinct primary intervention points within the sociotechnical stack, even as each requires multi-layer satisfaction in practice (Selbst et al. 2019). Epistemic humility primarily begins at the model layer: the system must be able to recognise when it is operating beyond its reliable scope (Wen, Howe, and Wang 2024; Griot et al. 2025). However, this capacity becomes meaningful for users only when it is made visible through interface affordances at the moment reliance is formed. Epistemic access primarily begins at the interface layer, through mechanisms for inspectability and contestability (Alfrink et al. 2023), but it depends on model and data infrastructures that can expose traceable reasoning, evidential grounding, and verifiable provenance. Resistance to epistemic injustice has the most distributed structure. It is shaped at the data layer by what forms of knowledge are represented and treated as credible (Kay, Kasirzadeh, and Mohamed 2024), at the model layer through ranking, retrieval, and credibility weighting, and at the interaction layer through whether user expertise is recognised as epistemically relevant. These intervention points indicate where evaluation can begin, while each condition must ultimately be assessed across the deployment as a whole. Surplus work at one layer cannot compensate for failure at another.

3.4 Non-Fungibility of the Three Conditions

We define epistemic humility (H), epistemic access (A), and resistance to epistemic injustice (R) as jointly necessary and non-fungible conditions of epistemic trustworthiness. By non-fungibility, we mean that the conditions are not mutually substitutable: performance above the required threshold on one condition cannot compensate for another falling below its contextually required threshold. This structure draws on plural, non-aggregative accounts of value, according to which distinct normative dimensions cannot be reduced to a single commensurable score (Anderson 1993). In the cases below, a condition's symbol indicates that its threshold is met, whereas its negation indicates that the condition falls below that threshold.

1. $H \wedge A \wedge \neg R$: **humility and access without resistance to epistemic injustice.** A system may communicate uncertainty clearly and make its outputs inspectable, yet rely on sources, categories, or credibility structures that marginalise particular users or communities (Kay, Kasirzadeh, and Mohamed 2024). In such cases, better calibration or more detailed explanations cannot correct whose knowledge is represented or treated as credible.
2. $H \wedge \neg A \wedge R$: **humility and resistance without access.** A system may signal its limitations and recognise users

as legitimate knowers, yet present outputs in forms that users cannot practically inspect, interpret, or contest. Effective oversight requires both access to relevant information and the practical capacity to intervene (Sterz et al. 2024). Recognition alone therefore does not enable users to evaluate or challenge the output.

3. $\neg H \wedge A \wedge R$: **access and resistance without humility.** A system may provide inspectable and contestable outputs while recognising users’ knowledge, yet communicate confidence that exceeds its evidential support (Xiong et al. 2024). Access and recognition cannot compensate for the absence of limitation signals needed to calibrate reliance.

These cases establish that the framework is minimal with respect to its constituent conditions: removing any condition would admit a user–system relation that satisfies the remaining two while retaining a distinct epistemic defect. Evaluation should therefore assess the conditions conjunctively. Rather than asking whether a system is trustworthy “overall”, evaluators should determine whether each condition meets its contextually required threshold and identify where failures arise, whom they affect, and with what consequences.

4 Applying the Framework: Case Analyses of Warranted Reliance

The following cases are selected to demonstrate the framework’s diagnostic value across different domains and failure modes. Each case meets four criteria: (1) the failure is publicly documented or empirically evaluated; (2) the system functions as an epistemic source, shaping users’ beliefs, judgements, or professional reasoning; (3) the reliance context is consequential, affecting legal, medical, employment, or institutional outcomes; and (4) the case exposes a failure of warranted reliance that is only partially captured by standard output-level evaluation, including accuracy metrics, fairness audits, safety scores, or citation display. We do not treat these cases as a representative sample; rather, we use them as theoretically informative cases that make different failures of warranted reliance analytically visible. Across these cases, warranted reliance fails through distinct epistemic pathways. The framework unifies these cases by treating them as failures of epistemic trustworthiness, while differentiating the specific condition foregrounded in each case.

4.1 Mata v. Avianca and Medical Metacognition: Epistemic Humility Failure

In *Mata v. Avianca*, two lawyers submitted a legal filing that cited non-existent cases generated by ChatGPT. The fabricated cases had plausible names, citations, quotations, and summaries. After their existence was questioned, one lawyer asked ChatGPT whether the cases were real, and the system reaffirmed that they were. The court sanctioned the lawyers and their firm for submitting fictitious opinions and continuing to defend them after opposing counsel and the court had raised doubts (*Mata v. Avianca, Inc.* 2023).

The case can be diagnosed as hallucination compounded by inadequate professional verification. This diagnosis captures the system’s generation of non-existent cases and the lawyers’ failure to verify them, but it leaves a further system-side question: did the output communicate that the cited cases were unverified, and did it preserve that limitation when the user requested confirmation? We analyse this requirement through epistemic humility, which requires systems to communicate material limitations in a form users can act on and to maintain those limitations across interaction unless new evidence warrants revision.

On this account, the case contains two humility failures. At generation, the system presented fabricated authorities without a limitation signal or direction to an authoritative verification source. At reconfirmation, it failed to revise, qualify, or abstain when explicitly asked whether the cases were real. An evaluation limited to the initial response would miss the second failure: improved first-turn calibration would not prevent a similar outcome if appropriate limitations disappeared when the user requested confirmation. This diagnosis does not diminish the lawyers’ professional responsibility to verify the authorities they submitted. It identifies an additional system-side failure within the interaction.

The *MetaMedQA* benchmark isolates this same second-order competence in a different domain, evaluating whether models can detect when correct answers are absent and reliably assess their own lack of knowledge (Griot et al. 2025). Its findings show that high first-order accuracy on standard medical evaluations does not necessarily translate into reliable metacognitive detection. We read this as direct evidence for the framework’s central claim about humility: the capacity to abstain or redirect is not a downstream consequence of better task performance, but a separately specifiable property that current evaluation regimes do not adequately target.

This diagnosis changes the remediation question. If humility is constitutive, the relevant interventions are neither solely model-level, meaning more accurate generation, nor solely user-level, meaning more careful verification. They are interactional: distinguishing generated from retrieved content, signalling uncertainty about citation existence, refusing to confirm claims the system cannot substantiate, and routing users to authoritative verification pathways.

Mata also shows why epistemic humility and epistemic access are jointly necessary rather than substitutable. The humility failure became consequential because users lacked an in-system verification pathway through which the system’s apparent certainty could have been contested before professional reliance was formed. At the same time, provenance alone would not be sufficient: if a system presents fabricated or weakly grounded claims with unwarranted confidence, users have little reason to activate verification mechanisms in the first place. The conditions therefore interact, but neither subsumes the other. Humility signals when reliance should be limited; access gives users the means to inspect the grounds on which the system is acting and to contest them.

Although resistance to epistemic injustice is not the primary failure mode in *Mata*, the case raises a related distributive question: who bears the burden of verification when sys-

tems fail to signal their limits? Professionals differ in their available time, expertise, institutional support, and access to authoritative databases, so when verification is shifted onto the user, humility and access failures may expose some users to greater professional risk than others. This is a distributive concern rather than a direct case of epistemic injustice; the next case turns to a failure in which epistemic standing itself is differentially assigned.

4.2 Resume Screening: Epistemic Injustice in Candidate Evaluation

Wilson and Caliskan audit three text-embedding models using 554 résumés and 571 job descriptions across nine occupations. They vary identity-signalling names while holding the remaining résumé information constant. Across 27 model–occupation comparisons, the models favour White-associated names in 85.1% of racial comparisons and male-associated names in 51.9% of gender comparisons. In the intersectional analysis, they favour White-male-associated names over Black-male-associated names in all 27 comparisons (Wilson and Caliskan 2024). These findings matter not only because the rankings are unequal, but also because ranking shapes visibility. In a retrieval-based screening process, a résumé’s position can affect whether and how prominently its evidence of qualification is presented to a recruiter. A qualification-irrelevant identity signal can therefore reduce the opportunity for a candidate’s qualifications to inform the decision.

This pattern is structurally analogous to Fricker’s account of pre-emptive testimonial injustice. Fricker defines testimonial injustice as a credibility deficit imposed on a speaker because of identity prejudice (Fricker 2007). Her account also includes cases in which prejudice prevents a person’s contribution from being sought or heard. Résumé retrieval is not itself a testimonial exchange, and our analysis attributes neither prejudice nor agency to the model. It nevertheless reproduces a structurally similar pattern: changing an identity-signalling name alters the likelihood that otherwise unchanged qualification information will enter the recruiter’s consideration. We use *visibility deficit* as a descriptive label for this algorithmically mediated form of pre-emptive exclusion. R1 therefore concerns not only how evidence is weighted after it is received, but also whether identity affects its consideration at all.

The benchmark establishes identity-sensitive retrieval but does not determine whether learned social associations, name frequency, or another representational mechanism produces it. Nor does it observe how recruiters interpret or rely on the rankings. It therefore supports an R1 diagnosis at the retrieval stage while leaving epistemic humility and access unassessed. Evaluating humility would require examining whether a deployed system communicates its known sensitivity to qualification-irrelevant identity signals. Evaluating access would require examining whether recruiters can inspect, test, and challenge that sensitivity and determine how much weight to assign the ranking. Such mechanisms could make the failure visible and contestable, but they would not correct the identity-contingent weighting itself. Remediation must also address the unjustified sensitivity in the data,

embeddings, or ranking procedure; until then, the ranking should not be treated as neutral evidence of candidate fit.

4.3 Legal RAG Systems: Epistemic Access Failure in Source-Grounded Tools

Retrieval augmentation is designed to improve factual grounding by connecting generated answers to authoritative sources (Fan et al. 2024; Lewis et al. 2020). In legal research, this design supports a plausible inference: an answer linked to accessible legal sources provides a basis for professional verification. We examine this inference using a preregistered evaluation of Lexis+ AI, Westlaw AI Assisted Research, and Ask Practical Law AI. Magesh et al. found that the tested versions produced hallucinated responses to between 17% and 33% of evaluated queries. Their analysis included both fabricated legal sources and misgrounded responses in which a real source did not support the proposition for which it was cited (Magesh et al. 2025).

Misgrounded citations show why source availability does not by itself establish epistemic access. A fabricated citation can be rejected once the user discovers that the cited case does not exist. A real citation that is misapplied requires the user to inspect the source, determine whether the cited passage states the court’s ruling or merely an observation that is not legally binding, assess its relevance and continuing legal validity, and compare it with the generated claim. Verification may therefore be possible in principle while remaining burdensome in practice. Citations can make an answer appear grounded even when determining whether that appearance is warranted requires substantial independent legal analysis (Magesh et al. 2025).

The benchmark provides direct evidence concerning verifiable linkage between claims and evidence (A1): the existence of a real citation does not show that the cited source supports the corresponding claim (Magesh et al. 2025). Under our framework, epistemic access also requires inspectable retrieval (A2). Users need sufficient information to assess why particular sources were selected and whether relevant qualifications, omissions, or conflicting sources affect the answer. Epistemic access further requires contestability (A3), through which users can challenge or correct an output when inspection reveals a problem. Magesh et al. do not directly evaluate A2 or A3. Determining whether these properties are satisfied would require additional evidence from the interfaces and professional workflows in which the tools are used. Our framework therefore adds a relational conclusion to the benchmark’s findings about outputs. Source grounding supports epistemic access only when users can verify the relationship between a claim and its evidence within realistic workflow constraints and act when that relationship fails. Evidence against an access diagnosis would include consistent support for each consequential proposition, successful verification by representative legal professionals within realistic time constraints, and an effective pathway for correction or escalation. If these requirements were met at contextually justified thresholds, occasional model errors would not by themselves establish a failure of epistemic access.

Misgrounded citations may also raise a question of epistemic humility if the system presents them with greater cer-

tainty than their evidential support warrants. The benchmark, however, does not directly assess how the systems communicate uncertainty or limitations. The case therefore provides direct evidence of failures in the linkage between claims and evidence, while the remaining properties of epistemic access and epistemic humility require separate evaluation.

4.4 IatroBench and the Identity-Contingent Refusal Pattern in Medical QA

While the preceding cases isolate individual conditions, IatroBench shows how a single sociotechnical failure can cut across several conditions. The preregistered benchmark examines whether safety-oriented model behaviour causes omission harm by withholding clinically relevant assistance. It evaluates 60 clinical scenarios across six frontier models, generating ten responses for each scenario–model pair and 3,600 responses in total. All responses were assessed for commission harm, caused by providing dangerous advice, and omission harm, caused by withholding necessary assistance. In a separate matched-framing analysis of 22 scenarios, the same clinical case was presented as either a patient’s question or a physician-framed consultation. Because the clinical facts remained unchanged, the analysis could test whether the framing of the user affected the assistance provided. Across the five models included in this comparison, physician framing produced more complete guidance than layperson framing. In the benchmark’s motivating example, a psychiatrist received detailed guidance on safely reducing a medication dose, whereas a patient facing the same clinical problem received a generic referral (Gringras 2026).

A conventional safety evaluation may interpret such referrals as appropriately cautious. IatroBench shows why this interpretation is insufficient. A standard LLM judge assigned an omission-harm score of zero to 73% of responses that a physician rated as harmful, with negligible agreement ($\kappa = 0.045$). In other words, physicians judged these refusals harmful because they withheld necessary assistance, whereas the automated evaluator judged nearly three-quarters of them harmless. Conventional LLM-based evaluation may therefore mistake withholding for safety. This discrepancy is especially consequential because the scenarios describe users for whom ordinary referral pathways are unavailable or have already failed. In these circumstances, a generic instruction to seek professional help may leave the user without a viable course of action while retaining the appearance of a safe response.

Our framework distinguishes three aspects of this pattern. Resistance to epistemic injustice is most directly implicated: holding the clinical facts constant while changing the speaker’s professional framing alters the assistance provided. Professional status therefore appears to affect whose need for guidance receives epistemic recognition. Epistemic access is also implicated because the patient-facing response does not reveal what information has been withheld, explain the basis of the refusal, or indicate how the refusal can be challenged. Finally, the refusals may also implicate epistemic humility. The same generic refusal may occur because the model lacks the relevant knowledge, because its

safety training causes it to withhold information it possesses, or because a downstream filter removes that information. This matters for epistemic humility because that condition requires the system to represent and communicate the limitations governing its response accurately. Because the response does not distinguish among these causes, users cannot determine whether the system is unable or unwilling to answer and cannot calibrate their reliance accordingly.

The benchmark thus documents role-contingent withholding, while our analysis shows why this pattern also raises concerns about epistemic access and humility. Confirming those diagnoses in deployment would require further evidence about refusal explanations, contestation mechanisms, and user understanding. The case nevertheless establishes a clear evaluative lesson: reliance on a refusal is not warranted merely because the refusal sounds cautious. Its basis must be communicated accurately and remain open to scrutiny, while any differentiation based on professional status must be explicit and epistemically justified.

5 Discussion

The framework developed in this paper centres a situated evaluative question: does a particular user have adequate grounds to rely on a particular output in context? In high-stakes GenAI use, the central question is not whether users accept system outputs, but whether they are positioned to accept, verify, contest, or reject those outputs on warranted grounds. This distinction matters because deference can arise through different pathways. Users may defer because accepting the output is the path of least resistance; because the system provides reliable, interpretable signals that make deference warranted; or because misleading reliability cues make unwarranted deference appear justified. We term these uninformed, informed, and misinformed deference. The last is especially damaging: users may engage carefully with expressed reliability signals and still be misled if those signals reflect artefacts of generation rather than reliable evidence of competence. Supporting epistemic trustworthiness therefore requires design and evaluation practices that preserve users’ capacity for warranted judgement, rather than merely increasing acceptance or apparent trust.

5.1 Tensions and Trade-offs

The three conditions also raise practical tensions for design and governance. First, epistemic humility may appear to reduce usefulness when uncertainty signalling or abstention is overused. This does not imply that systems should minimise uncertainty communication; rather, it motivates calibrated, context-specific humility rather than generic caution. Deployment evidence from AVA-AI, a domain-bounded multi-agent RAG system used by over 2,200 professionals across 116 countries, illustrates this tension in practice (Karnatak et al. 2026). In early deployment, when the curated corpus contained approximately 50 reports, abstention rates ranged from 40–70%; users interpreted refusals as evidence of limited system capability rather than as epistemic signals. After the corpus expanded to over 4,000 World Bank reports, abstention rates fell below 10%, and reasoned abstention be-

came legible as a reliability signal: the system was not refusing because it lacked capability, but because the available evidence did not warrant a claim. The lesson is not that systems should abstain less, but that abstention earns its epistemic function only when users can read it as a judgement about evidence rather than a limit of capability. Second, epistemic access may conflict with proprietary constraints when full provenance, reasoning traces, or model internals cannot be disclosed. In such cases, systems should not be represented as epistemically trustworthy for high-stakes use unless alternative audit mechanisms, contestation pathways, or evidence summaries are provided. Third, resistance to epistemic injustice is difficult to achieve at scale because epistemic recognition requires more than aggregate fairness metrics: it may require corpus audits, deployment-context validation, and engagement with affected communities whose knowledge is often underrepresented in dominant data regimes. Consequently, navigating these tensions while satisfying the three conditions requires deployment-specific design and governance.

5.2 Non-Fungibility and Its Implications for Evaluation

Our non-fungibility claim (Section 3.4) draws on the structural logic of plural, threshold-based normative frameworks, including the capabilities approach developed in welfare economics and political philosophy (Nussbaum 2011; Sen 1999). In these frameworks, a normative good cannot be reduced to a single averaged or aggregated score; rather, it requires securing multiple constitutive dimensions, each of which protects a distinct and irreducible aspect of that good (Anderson 1993). Analogously, epistemic trustworthiness should not be evaluated as the weighted sum or numerical average of epistemic humility, epistemic access, and resistance to epistemic injustice. Each condition protects against a structurally distinct pathway through which reliance can lose its epistemic warrant. Strong performance on one dimension therefore cannot compensate for serious failure on another. The conditions should instead be treated as thresholds that must each be satisfied to a level commensurate with the risks, users, and institutional consequences of the deployment context. This non-fungible structure has direct implications for AI evaluation methodology. Composite trustworthiness scores may be useful for summarising broad system performance, but they can obscure failures along individual dimensions. For example, a generative system may score highly on an aggregate rubric because it communicates uncertainty clearly and provides extensive retrieval provenance, while still failing to recognise a specific user group's standpoint as epistemically legitimate (Fricker 2007). In such cases, the aggregate score may suggest that the system is trustworthy overall, even though reliance remains unwarranted for particular users, communities, or institutional contexts.

To address this limitation, evaluators should use context-sensitive and layer-explicit audit methods that assess each relational condition independently as part of a diagnostic profile. Evaluations of epistemic humility should compare system-side uncertainty signals, retrieval quality, or evi-

dence coverage with the confidence communicated to users, including whether limitation-signalling is maintained across multi-turn interaction. Evaluations of epistemic access should assess whether practitioners can trace the evidence supporting a generative claim, determine whether cited sources actually support the generated text, and understand the output's operational boundaries without disproportionate cognitive or temporal effort. Evaluations of resistance to epistemic injustice should examine whether the pipeline discounts localised domain knowledge, marginalised testimony, or non-dominant language practices, and whether contestation mechanisms allow user expertise to enter the interaction as an active corrective resource.

5.3 Calibrated Friction as a Design Implication

A design implication of epistemic trustworthiness is that interfaces should not present all GenAI outputs as epistemically equivalent. When evidence is strong, uncertainty is low, and the stakes of use are limited, systems may support relatively fluent interaction. When evidence is sparse, conflicting, weakly grounded, or consequential for downstream action, the interface should introduce friction that prompts users to inspect, verify, or reconsider the output. We call this *calibrated friction*: interactional friction that increases in proportion to the epistemic risk of relying on a system output. The aim is not to make systems harder to use, but to prevent fluency from becoming unwarranted deference. Prior work shows that friction interventions, such as cognitive forcing functions that require engagement before acceptance, can reduce overreliance without eliminating appropriate reliance (Bućinca, Malaya, and Gajos 2021). We reframe this finding normatively: friction is not justified as a behavioural nudge but only when proportionate to genuine epistemic risk. Such friction may take several forms: claim-level uncertainty markers, prompts to inspect cited sources, flags for disagreement across retrieved evidence, or confirmation steps before acting on high-stakes outputs. Across these forms, friction should be grounded in epistemically meaningful reliability signals, such as retrieval quality, semantic uncertainty (Kuhn, Gal, and Farquhar 2023), or source agreement, rather than in generated self-assessments. The latter are especially risky because verbalised confidence often clusters at high certainty regardless of actual accuracy (Xiong et al. 2024), thereby reproducing the very miscalibration that friction is meant to counter. Calibrated friction surfaces a central design tension for epistemic agency. Too little friction encourages users to accept outputs that require verification; too much friction becomes paternalistic, obstructing reliance that may be justified. Standard friction interventions primarily address the first risk by slowing users down and prompting additional inspection. Yet this is insufficient in cases of misinformed deference, where users are already attending carefully to the system's expressed reliability signals. In such cases, the problem is not that the user is moving too quickly, but that the interaction directs attention towards the wrong epistemic cue. Friction that prompts further inspection of a poorly calibrated confidence score, a misleading citation, or a fluent explanation may therefore reinforce rather than correct the underlying failure. Calibrated

friction must therefore do more than interrupt action. It must redirect users towards cues that support epistemic assessment, rather than merely increasing attention to artefacts that may themselves be misleading. In this sense, calibrated friction operationalises epistemic humility by signalling when reliance should be limited, and epistemic access by creating conditions for users to inspect, question, or withhold reliance before acting. Its relationship to epistemic injustice is more limited but still important. Interface-level friction cannot substitute for data- and model-layer interventions that address representational harms, exclusion, or differential treatment. However, it can play a recognition-supporting role by making contingencies in system behaviour visible, for example by flagging when a response is sensitive to how the user is represented or how the query is framed. Properly designed, friction does not impede epistemic agency; it protects it by turning moments of uncertainty into opportunities for warranted judgement rather than occasions for misplaced deference.

5.4 Limitations

Our framework is normative and constitutive rather than implementation-prescriptive: it specifies the conditions an epistemically trustworthy system must satisfy, not the technical means by which they are realised. Mechanisms we discuss illustratively, such as reasoned abstention for humility, contestability for access, and participatory audit for resistance to epistemic injustice, are not fixed proxies or universal prescriptions; their adequacy depends on deployment context, users affected, and institutional stakes. The framework specifies what must be achieved; how it is achieved remains a local, collaborative design problem. Two limitations follow. First, resistance to epistemic injustice is the condition for which deployment-scale operationalisation is least developed: participatory audits and longitudinal studies of recognition patterns are well established in HCI but have not yet been systematically adapted to GenAI evaluation. Second, our cases are drawn from domains where harms are visible and documented; epistemic harms in less legible settings, where miscalibrated deference accumulates without producing discrete failures, are likely underrepresented and require further empirical investigation.

6 Conclusion

This paper establishes an account of when human reliance on generative AI is epistemically warranted rather than merely induced by fluency, framing, or interface cues. Grounded in social epistemology, we propose a user–system relational framework with three necessary, non-fungible conditions: epistemic humility, epistemic access, and resistance to epistemic injustice. Real-world cases across high-stakes domains show that hallucination, misgrounded citations, biased rankings, and identity-contingent refusals are not merely output defects but breakdowns in the relational conditions that make reliance warranted. While accuracy, calibration, safety, and fairness remain necessary, they do not capture the interactional conditions under which professional deference forms. We draw two implications. First,

evaluation must treat humility, access, and resistance to epistemic injustice as a diagnostic profile showing where and for whom reliance fails, rather than as a composite score. Second, design must support warranted judgement as reliance forms by introducing calibrated friction that is proportionate to epistemic risk and directs users to key reliability cues rather than generated self-assessments. Accordingly, responsible GenAI deployment requires treating these conditions as constitutive requirements and building systems that give users the grounds to inspect, contest, withhold, or appropriately extend reliance on AI outputs.

References

- Alfrink, K.; Keller, I.; Kortuem, G.; and Doorn, N. 2023. Contestable AI by Design: Towards a Framework. *Minds and Machines*, 33(4): 613–639.
- Anderl, C.; Klein, S. H.; Sarigül, B.; Schneider, F. M.; Han, J.; Fiedler, P. L.; and Utz, S. 2024. Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Sci. Rep.*, 14(1): 17127.
- Anderson, E. 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.
- Appel, R.; McCrory, P.; Tamkin, A.; McCain, M.; Neylon, T.; and Stern, M. 2025. Anthropic Economic Index report: Uneven geographic and enterprise AI adoption. arXiv:2511.15080.
- Baier, A. 1986. Trust and Antitrust. *Ethics*, 96(2): 231–260.
- Bommasani, R.; Liang, P.; and Lee, T. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1): 140–146.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.
- Cau, F. M.; and Spano, L. D. 2025. Exploring the Impact of Explainable AI and Cognitive Capabilities on Users’ Decisions. arXiv:2505.01192.
- Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).
- Chiesurin, S.; Dimakopoulos, D.; Sobrevilla Cabezudo, M. A.; Eshghi, A.; Papaioannou, I.; Rieser, V.; and Konstas, I. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 947–959. Toronto, Canada: Association for Computational Linguistics.
- Craig, E. 1990. *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*. Oxford: Clarendon Press.
- Dell’Acqua, F.; McFowland, E., III; Mollick, E.; Lifshitz, H.; Kellogg, K. C.; Rajendran, S.; Kraye, L.; Candelon, F.; and Lakhani, K. R. 2026. Navigating the jagged technological frontier: Field experimental evidence of the effects of

- artificial intelligence on knowledge worker productivity and quality. *Organ. Sci.*, 37(2): 403–423.
- Dennett, D. C. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
- Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1): 114–126.
- DiMarco, M. 2023. Cooperative Epistemic Trustworthiness. *Philosophy of Science*, 90(5): 1100–1109.
- Dotson, K. 2011. Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, 26(2): 236–257.
- Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, 6491–6501. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704901.
- Ferrario, A. 2024. Justifying our credences in the trustworthiness of AI systems: A reliabilistic approach. *Science and Engineering Ethics*, 30(6): 55.
- Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
- Gringras, D. 2026. IatroBench: Pre-Registered Evidence of Iatrogenic Harm from AI Safety Measures. arXiv:2604.07709.
- Griot, M.; Hemptinne, C.; Vanderdonckt, J.; and Yuksel, D. 2025. Large Language Models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1): 642.
- Heersmink, R.; De Rooij, B.; Clavel Vázquez, M. J.; and Colombo, M. 2024. A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, 26(3): 41.
- Hila, A. 2026. The epistemological consequences of large language models: rethinking collective intelligence and institutional knowledge. *AI & SOCIETY*, 41(1): 79–97.
- Jonas, E.; Greussing, E.; and Taddicken, M. 2025. Disentangling (Hybrid) Trustworthiness of Communicative Generative AI as Intermediary for Science-Related Information—Results from a Qualitative Interview Study. *Human-Machine Communication*, 11: 213–236.
- Karnatak, N.; Chatila, M.; Pinzón Hernández, D. A.; Yazdanfar, R.; Dugas, M.; and Vakis, R. 2026. Learning from AVA: Early Lessons from a Curated and Trustworthy Generative AI for Policy and Development Research. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, CHI '26*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic Injustice in Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 684–697.
- Koskinen, I. 2024. We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, 38(4): 458–475.
- Kowald, D.; Scher, S.; Pammer-Schindler, V.; Müllner, P.; Waxnegger, K.; Demelius, L.; Fessler, A.; Toller, M.; Mendoza Estrada, I. G.; Šimić, I.; Sabol, V.; Trügler, A.; Veas, E.; Kern, R.; Nad, T.; and Kopeinik, S. 2024. Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data*, 7: 1467222.
- Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.
- Lackey, J. 2008. *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.
- Lange, B. 2026. Epistemic Deference to AI. In *Bridging the Gap Between AI and Reality: AISoLA 2024*, volume 16032 of *Lecture Notes in Computer Science*, 174–186. Springer.
- Lee, J. D.; and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1): 50–80.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- Liao, Q.; and Sundar, S. S. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, 1257–1268. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; and Ho, D. E. 2025. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 22(2): 216–242.
- Marchal, N.; Chan, S.; Franklin, M.; Revel, M.; Keeling, G.; Fischli, R.; Chandra, B.; and Gabriel, I. 2026. Architecting Trust in Artificial Epistemic Agents. arXiv:2603.02960.
- Mata v. Avianca, Inc. 2023. Opinion and Order on Sanctions, 678 F. Supp. 3d 443 (S.D.N.Y.). Case No. 22-cv-1461 (PKC), United States District Court, Southern District of New York. Decided June 22, 2023.
- Medina, J. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press.
- Mehrotra, S.; Degachi, C.; Vereschak, O.; Jonker, C. M.; and Tielman, M. L. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.*, 1(4).
- Metzger, L.; Miller, L.; Baumann, M.; and Kraus, J. 2024. Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.
- Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

- Morris, M. R.; Altman, D.; Belfield, H.; Goemans, A.; Iqbal, H.; Burnell, R.; Gabriel, I.; Albanie, S.; and Dafoe, A. 2026. Characterizing Jaggedness Aids Safety & Usability. Technical report, Google DeepMind.
- Nussbaum, M. C. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Belknap Press of Harvard University Press.
- Parasuraman, R.; and Manzey, D. H. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3): 381–410.
- Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Sanderson, C.; Douglas, D.; Lu, Q.; Schleiger, E.; Whittle, J.; Lacey, J.; Newnham, G.; Hajkowicz, S.; Robinson, C.; and Hansen, D. 2023. AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*, 4(2): 171–187.
- Sanderson, C.; Douglas, D. M.; and Lu, Q. 2023. Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–7.
- Scheman, N. 2001. Epistemology Resuscitated: Objectivity as Trustworthiness. In Tuana, N.; and Morgen, S., eds., *Engendering Rationalities*, 23–52. Albany, NY: State University of New York Press.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Sen, A. 1999. *Development as Freedom*. Oxford: Oxford University Press.
- Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S. M.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.
- Simhi, A.; Itzhak, I.; Barez, F.; Stanovsky, G.; and Belinkov, Y. 2025. Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 14665–14688. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.
- Simion, M.; and Kelp, C. 2023. Trustworthy Artificial Intelligence. *Asian Journal of Philosophy*, 2(1): 8.
- Simon, J. 2026. Generative AI, Quadruple Deception & Trust. *Social Epistemology*, 40(1): 101–115.
- Sivaraman, V.; Bukowski, L. A.; Levin, J.; Kahn, J. M.; and Perer, A. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.
- Slome, E. 2026. Decision Ownership and Deference to Healthcare AI. *Journal of Bioethical Inquiry*, 1–6.
- Sosa, E. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford: Oxford University Press.
- Sterz, S.; Baum, K.; Biewer, S.; Hermanns, H.; Lauber-Rönsberg, A.; Meinel, P.; and Langer, M. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 2495–2507. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Tanchuk, N. J.; and Taylor, R. M. 2025. Personalized Learning with AI Tutors: Assessing and Advancing Epistemic Trustworthiness. *Educational Theory*, 75(2): 327–353.
- Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M. S.; and Krishna, R. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).
- Wen, B.; Howe, B.; and Wang, L. L. 2024. Characterizing LLM Abstention Behavior in Science QA with Context Perturbations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 3437–3450. Miami, Florida, USA: Association for Computational Linguistics.
- Whittlestone, J.; Nyrup, R.; Alexandrova, A.; and Cave, S. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 195–200. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Wilson, K.; and Caliskan, A. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 1578–1590.
- Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.
- Yin, M.; Wortman Vaughan, J.; and 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*, 1–12.
- Yun, B.; Feng, D.; Chen, A. S.; Nikzad, A.; and Salehi, N. 2025. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 295–305.