



EPIC-PP-2026-001 · RESEARCH ARTICLE · AI GOVERNANCE

Borrowed Certainty: The AI Risk That Feels Like Relief

Why an accurate machine can leave us less safe, and a vigilance-preservation criterion for AI ethics

EPIC – Emirati Perspectives and Insight Centre

Seeing what others miss...

20 October 2026

CC BY 4.0 · Creative Commons Attribution 4.0 International · creativecommons.org/licenses/by/4.0

Borrowed Certainty: The AI Risk That Feels Like Relief

EPIC – EMIRATI PERSPECTIVES AND INSIGHT CENTRE

EPIC-PP-2026-001 · AI Governance · Published 20 October 2026

© 2026 EPIC – Emirati Perspectives and Insight Centre. Released under Creative Commons Attribution 4.0 International (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>

EPIC is an independent research initiative in formation. It is not yet a registered or licensed entity. EPIC does not accept donations, payments or contracts.

Abstract

The prevailing picture of AI risk is a picture of error: the machine is wrong and we are misled. This paper argues that a more insidious risk lies in relief. A fluent, confident answer relieves the discomfort of not knowing, and that relief tends to switch off the vigilance that would have caught the machine's mistakes. The result is borrowed certainty: the psychological state of certainty without the epistemic achievement that would make it knowledge. Drawing on Sosa's virtue epistemology, Kant's account of self-incurred minority, Fromm's analysis of the flight from freedom and the philosophy of epistemic dependence and trust, the paper develops four concepts: borrowed certainty; the relief threshold, at which confidence cues end inquiry; the distinction between waking and sleeping trust; and a vigilance-preservation criterion, under which AI systems are judged by what they do to their users' vigilance and not only by their accuracy. A simple expected-harm model shows when a highly accurate system that extinguishes checking produces more undetected harm than a less accurate one that keeps users alert. Published work on automation bias and over-reliance on large language models is consistent with this picture. Taha Abdurrahman's trusteeship ethics and Ibn Rushd's account of demonstration ground the claim that responsibility cannot be transferred to a machine. Design and policy implications and an empirical agenda follow.

Keywords: Borrowed certainty; Epistemic trust; Automation bias; Over-reliance on large language models; Virtue epistemology; Trusteeship ethics (i'timāniyya)

1 Introduction

We usually picture the danger of artificial intelligence as error. The machine invents a court case, misreads a scan, miscalculates a dose, and someone is harmed. The picture is not false, but it is incomplete, because it places the whole of the risk inside the machine.

This paper argues that an overlooked part of the risk sits inside us, and that it does not feel like danger at all. It feels like relief.

Consider what happens at the moment a fluent answer arrives. Before it there was a question, and with the question a small discomfort: the tension of not knowing, an open loop that the mind would like to close. The answer closes it. It is grammatical, well organised and unhesitating. The discomfort dissolves, and with it the urge to keep looking. That dissolution is pleasant. It is also an epistemic event, because it marks the point at which inquiry stops. If the answer was right, little is lost beyond a chance to learn. If it was wrong, the error now passes through a mind that has stopped guarding the door.

We call the resulting state *borrowed certainty*: the psychological state of certainty without the epistemic achievement that would make it knowledge. Fluent relief feels like knowledge. It is not knowledge. It is certainty on loan, and the lender is not answerable for what we do with it.

The thesis of the paper is that the overlooked risk of AI is not error but relief, and that relief matters because it switches off vigilance. From this follows a claim that sounds paradoxical and is not: a system that is right ninety-nine times in a hundred but extinguishes checking can, under conditions we specify, be more dangerous than a system that is right ninety times in a hundred and keeps its users awake. Accuracy belongs to the machine. Vigilance belongs to the human. Harm is a joint product of the two, and an ethics that measures only the first is measuring the wrong thing.

The paper is not a literature review. It proceeds by conceptual analysis, a deliberately simple formal model, a check of that model against published empirical findings, and a reading of two traditions rarely placed side by side in the philosophy of technology: the Enlightenment's worry about intellectual minority, and the Arabic-Islamic idea of the trust (*amāna*) that a human being carries and cannot hand over. It coins and defends four concepts: *borrowed certainty*; the *relief threshold*, the point at which confidence cues end inquiry whether or not the source deserves it; the distinction between *waking trust* and *sleeping trust*; and a *vigilance-preservation criterion* for judging AI systems by what they do to their users' vigilance. It also states formally the conditions under which gains in accuracy defeat themselves.

We are not the first to use the word vigilance here. Sperber and colleagues (2010) argued that communication is sustained by cognitive mechanisms of *epistemic vigilance*; their concern is descriptive and evolutionary, ours normative: vigilance as what a technology should be evaluated for preserving. Nor is over-reliance on automation a new discovery (Parasuraman & Riley, 1997; Parasuraman & Manzey, 2010). What we add is an account of why fluent, conversational systems are a distinctive case, since they act on the felt sense of having understood and not merely on the perceived reliability of a tool, and a philosophical account of what is lost when vigilance falls asleep.

Sections 2 and 3 develop the epistemology and psychology of relief. Section 4 reviews the evidence, Section 5 builds the model, Section 6 develops the argument from trusteeship, Section 7 answers objections, and Sections 8 and 9 draw implications for design, policy and research.

2 Certainty is not knowledge

2.1 A state and an achievement

Certainty names a psychological state: felt closure, an absence of doubt, a readiness to act without further question. *Knowledge* names an achievement: a success creditable to the knower. The two often travel together, which is why we confuse them, but they come apart in both directions. The anxious student who answers correctly knows without feeling certain; everyone who has been confidently wrong has felt certain without knowing.

Ernest Sosa's virtue epistemology gives the distinction a precise shape. Sosa treats belief as a kind of performance aimed at truth, and evaluates performances on three dimensions. A performance is *accurate* if it succeeds, *adroit* if it manifests competence, and *apt* if it is accurate *because* adroit (Sosa, 2007, 2011). His recurring example is the archer. A shot can hit the target through a lucky gust of wind, in which case it is accurate but not apt; it can be skilful and miss; or it can hit the target because of the archer's skill. Knowledge, at the basic level, is apt belief. At a higher level, which Sosa calls knowing full well, the knower's first-order aptness is itself guided by a second-order competence: a sensitivity to risk, to when it is appropriate to believe and when to withhold (Sosa, 2011).

Notice where certainty sits in this scheme. Nowhere. Certainty is not one of the dimensions of epistemic evaluation; it is a feeling that can accompany an apt belief, an inapt but true belief, or a false belief, and it can feel identical in all three cases. That is the core of the difficulty. The feeling of having arrived is the same whether one arrived by skill, by luck or not at all.

2.2 The parcel and the wrong address

An image helps. Think of certainty as a parcel. Ordinarily it is delivered to someone who has done the work, someone who has checked, compared, reasoned and earned the closure the parcel contains. In that case the address is right: the certainty belongs to the person who receives it. When a fluent machine answers my question, the parcel still arrives. It is real, the package is solid, and I sign for it. But it has been delivered to the wrong address. The certainty was meant for a mind that did the epistemic work, and it has landed in a mind that did not. The feeling is genuine; the entitlement is not.

The image should not mislead. Borrowed certainty is not simply belief based on what someone else says. Hardwig (1985) argued that where an expert is better placed than I am, rationality may require me to defer rather than think the matter through myself, and Goldberg (2010) developed a social reliabilism on which the reliability that justifies a testimonial belief can extend to the speaker's cognitive processes. Much of what I know I know on loan, and none the worse for it.

What is distinctive about *borrowed* certainty lies in the hearer's contribution. Even deferential testimonial belief requires a competence in relying: sensitivity to defeaters, noticing when an answer does not fit, some grip on when a source is to be trusted, and knowing when the stakes call for a second look. This is the testimonial analogue of Sosa's second-order competence. It is precisely this competence that relief switches off.

2.3 A definition

We propose the following.

A person S holds a *borrowed certainty* that p when (i) S is certain, or nearly so, that p; (ii) S's certainty is produced chiefly by presentational features of a source's output (its fluency, assurance, completeness and tone) rather than by S's exercised competence in assessing that source; and (iii) the same cues that produced the certainty have disengaged S's dispositions to check, query or withhold.

Three consequences follow. First, a borrowed certainty may be true. The point is not that it is false but that its truth, when it is true, is not creditable to the believer's handling of the source. The luck involved is not in the world, as in the archer's gust of wind, but in the reliance itself. Second, a borrowed certainty is brittle. Because the believer did not exercise the competence that would let them defend, qualify or revise the belief, they cannot do those things when challenged. Third, and most important, borrowed certainty is self-concealing. Condition (iii) means that the very state which would let S notice the problem has been switched off by the cue that created it.

2.4 Why fluency, and why now

Why should presentational features produce certainty at all? Research on processing fluency shows that the subjective ease of processing feeds into many judgements, including judgements of truth and confidence (Alter & Oppenheimer, 2009). Nguyen (2021) gives the diagnosis a sharp philosophical form. On his account we use the feeling of clarity "as a signal that we have investigated a matter sufficiently", so that "the sense of clarity functions as a thought-terminating heuristic". Those who understand this can exploit it and "induce us to terminate our inquiries too early" (Nguyen, 2021, abstract). Nguyen was not writing about AI, but his analysis applies with special force to language models, whose outputs are fluent by construction: generators of seductive clarity at scale.

The Popperian tradition supplies the contrast. For Popper (1963/2002), knowledge grows through conjecture and refutation, and even our best theories are to be held open to criticism; the mark of knowledge is the critical attitude, not certainty. Borrowed certainty converts a conjecture into a closed matter without its ever having faced a test.

3 The psychology of relief

3.1 Why not-knowing is uncomfortable

Why do we want certainty so badly? Kruglanski and Webster (1996) described a *need for cognitive closure*, a desire for a definite answer and an aversion to ambiguity, which drives *seizing* on early information and *freezing* on the answer once reached, and which varies between people and situations. An interface that delivers a definite answer instantly is perfectly shaped to satisfy that need.

We add a conjecture. If the discomfort of not knowing is aversive, a confident answer is not just informative but relieving, and the relief may reinforce the habit of asking and accepting without checking. No evidence reviewed below directly measures this affective pathway; Section 9 proposes how to.

3.2 Fromm: the soft authority of the interface

Erich Fromm's *Escape from Freedom* (1941) is usually read as a book about the politics of the 1930s, but its argument is psychological. Modern freedom released individuals from older bonds and left them isolated, anxious and powerless. People escape that anxiety through characteristic mechanisms, which Fromm called authoritarianism, destructiveness and automaton conformity: submitting to a power outside the self, attacking the world that threatens, or becoming what everyone else expects.

Fromm's escapes ran towards persons and crowds. The authority of his century demanded obedience, loyalty and belief. The authority of ours is softer and, for that reason, harder to resist. It is an interface, and it asks for almost nothing: not obedience, not loyalty, not even belief in any strong sense. It asks only that we stop checking. The anxiety Fromm described, the burden of having to judge for oneself in an uncertain world, is real, and the conversational machine offers a way to set it down without any of the indignity of submission. One does not kneel to a chatbot. One simply stops looking.

3.3 Kant: the book that understands for me

Kant's answer to the question "What is enlightenment?" begins with a definition: "Enlightenment is the human being's emergence from his self-incurred minority" (Kant, 1784/1996, AA 8:35). Minority is the inability to use one's own understanding without direction from another, and it is self-incurred when its cause is not a lack of understanding

but a lack of resolve and courage. Hence the motto, *Sapere aude!*, have the courage to use your own understanding. Kant is unsparing about why people remain minors: "It is because of laziness and cowardice". And then the passage that might have been written for our moment:

It is so comfortable to be a minor! If I have a book that understands for me, a spiritual advisor who has a conscience for me, a doctor who decides upon a regimen for me, and so forth, I need not trouble myself at all. I need not think, if only I can pay; others will readily undertake the irksome business for me. (Kant, 1784/1996, AA 8:35)

A conversational AI is the book that understands for me, made interactive, personalised, instant and nearly free. Kant's minors had to find the book and pay for it; that friction limited how far minority could extend. The limit is gone. AI does not invent intellectual immaturity. It makes immaturity convenient at scale.

Kant is careful, and we should be too. His target is not the book but the book *that understands for me*. Reading is not minority; letting the reading replace my understanding is. Kant also warns that "precepts and formulas ... are the ball and chain of an everlasting minority" (AA 8:35–36): not instruments as such, but instruments that substitute for one's own reason. The same distinction will reappear below as the difference between waking and sleeping trust.

3.4 Relief as the hinge

Fromm and Kant name different motives, anxiety in one case and laziness and cowardice in the other. The confident answer satisfies both at once, removing the effort of inquiry and the discomfort of not knowing in a single stroke. That is why relief, rather than error, is the hinge. Error is a property of an output. Relief is an event in a person, and it is the event at which vigilance is lost.

There is also an asymmetry in what we notice. Relief is felt; lost vigilance is not. We are aware of the pleasant closure but not of the alarm that no longer sounds. The risk is invisible to the person who bears it, which is why it cannot be left to individual users alone.

3.5 The relief threshold

We can now name the point at which this happens. Inquiry continues while felt uncertainty remains above some threshold, and stops when it falls below it. Call this the *relief threshold*. Confidence cues, such as fluency, assurance, an absence of hedges and a tidy structure, lower felt uncertainty. When they push it below the threshold, inquiry ends.

The threshold responds to *felt* uncertainty, which tracks cues rather than the source's actual reliability, so a system can cross it whether or not it deserves to. And the threshold moves: plausibly with stakes, time pressure, fatigue, need for closure and self-confi-

dence, which survey evidence links to critical engagement with AI output (Lee et al., 2025). A system therefore has a *relief profile*: how reliably, and on which questions, it pushes users below their threshold.

4 Evidence from automation bias and LLM over-reliance

Is the argument consistent with what we know empirically? We select findings that bear directly on the mechanism and report only results checked against the original sources. A caveat: no study we know of has directly measured relief as a mediator of reduced checking. What the evidence shows is that the components of the mechanism are real.

4.1 Automation bias is robust and is not cured by expertise

Long before LLMs, human-factors research established that people over-rely on automated aids. Parasuraman and Riley (1997) named over-reliance *misuse*. Bainbridge (1983) identified the "ironies of automation": automating a task turns the human into a monitor, a role people perform poorly, so that the more advanced the automation, the more crucial and the more difficult the human contribution becomes.

Skitka, Mosier and Burdick (1999) gave the phenomenon its now standard form. In a simulated flight task, participants without an automated aid outperformed those who had a very reliable, but not perfectly reliable, aid on the monitoring task. Participants with the aid made *omission errors*, failing to respond to events the aid did not flag, and *commission errors*, following the aid's directions even when these contradicted their training and other, fully valid, indicators. Among the explanations the authors discussed were economy of cognitive effort, diffusion of responsibility and the treatment of automation as an authority.

Parasuraman and Manzey's (2010) review found that complacency emerges especially under multiple-task load, occurs in naive and expert users alike and is not overcome by simple practice, and that automation bias cannot be prevented by training or instructions alone and affects teams as well as individuals. Attention is central to both. Goddard, Roudsari and Wyatt's (2012) systematic review of automation bias in clinical decision support (74 studies included from 13,821 screened) identified trust and confidence in the system, workload, task complexity and time pressure as mediators, and listed training, emphasising user accountability, the placement of advice on screen, displaying updated confidence levels and providing information rather than recommendations as candidate mitigators. Lee and See (2004) framed the design goal as *appropriate reliance*: trust calibrated to the system's actual capabilities.

Over-reliance, in short, is not a failure of the ignorant but a general feature of attention under reliable assistance.

4.2 Better aids can produce less effort

The most direct evidence for the accuracy–vigilance trade-off comes from a preregistered field experiment by Dell'Acqua (2022), whose title, "Falling asleep at the wheel", names the phenomenon. Some 181 professional recruiters each evaluated 44 résumés with the help of an algorithm of either higher or lower quality. Recruiters given the higher-quality AI were *less* accurate in their assessments than those given the lower-quality AI. Those with the weaker tool spent more time and effort, followed the algorithm less automatically, and improved over time; the effects were driven by experienced recruiters. Dell'Acqua's accompanying model formalises the intuition that as AI quality rises, humans have less incentive to remain attentive. The paper is a working paper and should be read with that status in mind, but its design is strong and its result is exactly what the vigilance-preservation argument predicts.

4.3 Fluent wrongness persuades

Dell'Acqua and colleagues (2026) studied 758 consultants at Boston Consulting Group working on realistic tasks with and without access to a large language model. On tasks inside what the authors call the "jagged technological frontier" of AI capability, consultants with AI completed 12.2% more tasks, 25.1% more quickly and with markedly higher quality. On a task deliberately chosen to lie outside the frontier, the picture reversed: 84.5% of the control group reached the correct answer, against 60% and 70.6% in the two AI conditions. The finding that matters most here is that AI-assisted answers were rated as more coherent and persuasive *even when wrong*. Borrowed certainty does not stay with the borrower. It is passed on, in improved form, to the next reader, who has even less access to its origins.

4.4 Confidence cues are decoupled from reliability

Two studies show that the cues users read as confidence do not track whether the model is right. Zhou, Hwang, Ren and Sap (2024) found that language models are reluctant to express uncertainty and that, when asked to express confidence, they are frequently overconfident, with high error rates (an average of 47%) among confident responses. In user studies, people relied on plain, unmarked statements nearly 90% of the time: a plain statement is read as a confident one. Overconfident miscalibration had lasting effects on later reliance. Strikingly, the authors traced part of the problem to the alignment pipeline itself: preference data and reward models showed a bias against expressions of uncertainty.

Steyvers and colleagues (2025) identified a *calibration gap* and a *discrimination gap* between what models "know" and what people think they know. With default explanations, participants overestimated the accuracy of LLM answers, and their ability to tell correct from incorrect answers was much weaker than the model's own internal confidence

would allow (an area under the curve of roughly 0.59–0.60 for humans against roughly 0.75–0.78 for the models' internal confidence). Longer explanations raised user confidence without improving discrimination. Adjusting the explanations so that their uncertainty language matched the model's internal confidence narrowed both gaps.

Together these studies establish the precondition of borrowed certainty: the features that produce certainty in users are not reliable signals of correctness, and some, such as length and the absence of hedges, actively mislead.

4.5 Uncertainty and friction help, but only partly

Can contrary cues wake vigilance? Partly. In a preregistered experiment (N = 404) on medical questions, Kim and colleagues (2024) gave participants an LLM-based system that was correct on half of them. Without uncertainty expressions, access to the system *lowered* accuracy compared with no access at all (63.9% against 74.2%). When the system expressed uncertainty in the first person ("I'm not sure, but..."), accuracy rose to 72.8%. On the questions where the system was wrong, participants without the system answered correctly 64.7% of the time, those with the unhedged system only 33.0%, and those with first-person hedges 47.8%. Only 19.1% of participants with the unhedged system reported searching for themselves, against 92.9% without it. Hedging reduced over-reliance but did not eliminate it.

Buçinca, Malaya and Gajos (2021) found that *cognitive forcing functions*, such as asking people to decide before seeing the AI's recommendation, reduced over-reliance compared with simple explanations; but the most effective designs were the least liked, and benefits were greater for people high in need for cognition. Vasconcelos and colleagues (2023), across five studies (N = 731), showed that over-reliance can be understood as a strategic choice: people engage with AI explanations when the benefit of doing so outweighs the cost, and explanations reduce over-reliance when they lower the cost of verification relative to the difficulty of the task.

Vigilance, then, responds to design, but only partly, and the designs that preserve it best cost something in satisfaction. That is the source of the strongest objection, considered in Section 7.

4.6 Offloading in the wild

Survey evidence is weaker in design but closer to everyday use. Lee and colleagues (2025) surveyed 319 knowledge workers, who provided 936 examples of using generative AI at work. Higher confidence in generative AI was associated with less critical thinking, while higher self-confidence was associated with more. Critical effort shifted towards verification, integration and task stewardship, and was reported to fall for routine, low-stakes tasks and under time pressure. Gerlich (2025), in a mixed-methods study of 666 participants, found a significant negative correlation between frequent AI tool use and

critical-thinking scores, mediated by cognitive offloading, with younger participants showing greater dependence. Both are correlational and rely substantially on self-report (and the latter has a published correction to one table, with conclusions stated to be unaffected); they are consistent with the mechanism, not proof of it. Messeri and Crockett (2024) have likewise warned that AI tools in science risk producing illusions of understanding, in which researchers believe they understand more than they do.

4.7 What the evidence does and does not show

Two strands converge: confidence cues are decoupled from reliability (4.3–4.4), and vigilance falls as perceived reliability rises, even among experts and even at a cost to joint performance (4.1–4.2). Their conjunction is what borrowed certainty describes. The evidence does not yet show that felt relief is the mediator, or where the relief threshold lies; Section 9 frames these as hypotheses.

5 Borrowed certainty: a conceptual model

5.1 Components

The argument can be stated with four quantities for a given class of decisions:

- e , the system's error rate: the probability that an output is wrong;
- c , the probability that the user checks an output *when it is wrong*;
- d , the probability that a check, once made, detects the error;
- S , the stakes: the harm caused by an undetected error.

In its simplest form, which assumes that every check catches the error ($d = 1$), the expected harm per decision is:

$$H = e \times (1 - c) \times S$$

The general form allows for imperfect checking:

$$H = e \times (1 - c \cdot d) \times S$$

The model is deliberately austere: it ignores human error on correct outputs and treats stakes as uniform within a class. Its purpose is to make one structural fact visible. The error that matters is not the error the system makes but the error that survives the human. Accuracy enters the equation once; vigilance multiplies every error.

5.2 A worked illustration

The following numbers are *invented for illustration only*. They are not estimates of any real system, population or domain.

Let System A be right 99% of the time ($e = 0.01$) and relieving enough that users check only 5% of outputs ($c = 0.05$). Let System B be right only 90% of the time ($e = 0.10$) but presented in a way that keeps users checking 95% of outputs ($c = 0.95$). With perfect detection and equal stakes, per 10,000 decisions:

- System A: $0.01 \times (1 - 0.05) = 0.0095$, or 95 undetected errors;
- System B: $0.10 \times (1 - 0.95) = 0.0050$, or 50 undetected errors.

System A makes ten times fewer errors but lets through almost twice as many. This is the sense in which a 99%-accurate system that extinguishes checking can be more dangerous than a 90%-accurate one that keeps users alert.

Honesty requires showing where the illustration breaks down. Table 1 varies the assumptions.

Table 1. Illustrative comparison of undetected errors (invented parameters, not empirical estimates)

Scenario	System A (99% accurate)	System B (90% accurate)	Undetected harm per 10,000 decisions (A / B)	A ÷ B
1. Perfect detection	$c = 0.05, d = 1, S = 1$	$c = 0.95, d = 1, S = 1$	95 / 50	1.90
2. Near-perfect detection	$c = 0.05, d = 0.95, S = 1$	$c = 0.95, d = 0.95, S = 1$	95.2 / 97.5	0.98
3. Imperfect detection	$c = 0.05, d = 0.80, S = 1$	$c = 0.95, d = 0.80, S = 1$	96 / 240	0.40
4. Stakes creep	$c = 0.05, d = 0.90, S = 3$	$c = 0.90, d = 0.90, S = 1$	286.5 / 190 (stakes-weighted)	1.51
5. Vigilance-preserving A	$c = 0.60, d = 0.90, S = 1$	$c = 0.90, d = 0.90, S = 1$	46 / 190	0.24

Note. ILLUSTRATIVE ONLY. All parameter values are hypothetical and chosen to show how the model $H = e \times (1 - c \cdot d) \times S$ behaves; they are not drawn from any study or system. $e_A = 0.01$, $e_B = 0.10$ throughout. Values above 1 in the last column mean the more accurate system lets more harm through.

Scenario 3 matters as much as Scenario 1. When checking is imperfect, the less accurate system loses its advantage, because each of its errors may slip past even an alert user. The paradox is not a law that less accurate systems are safer. It is a conditional claim: accuracy gains can be outweighed by vigilance losses when the vigilance lost was frequent and effective. Scenario 4 adds a mechanism we suspect is common, *stakes creep*: a

trusted system gets used for weightier decisions, so its rare undetected errors cost more. Scenario 5 shows the point of the exercise. The best outcome comes not from a less accurate system but from an accurate one that keeps users checking *where its errors are*. The vigilance-preservation criterion is not opposed to accuracy. It asks for accuracy and vigilance together.

5.3 The break-even condition

The comparison generalises. System B produces less undetected harm than System A, at equal stakes, exactly when

$$e_B / e_A < (1 - c_A \cdot d) / (1 - c_B \cdot d)$$

Call the right-hand side the *vigilance ratio*. A more accurate system is safer only if its accuracy advantage is larger than the vigilance ratio. In Scenario 1 the accuracy advantage is 10 and the vigilance ratio is 19, so the less accurate system wins. Engineering effort is usually directed at the left-hand side of this inequality. The argument of this paper is that the right-hand side is equally real and almost never measured.

5.4 When accuracy gains defeat themselves

Checking is not fixed. It depends on displayed confidence, track record, stakes and the user's relief threshold. Two propositions summarise the evidence; a third follows.

Proposition 1 (relief effect). Holding accuracy fixed, stronger confidence cues lower the probability of checking. The studies of Zhou et al. (2024), Steyvers et al. (2025) and Kim et al. (2024) are consistent with this.

Proposition 2 (complacency effect). Holding cues fixed, a better track record lowers the probability of checking. Dell'Acqua (2022) and the automation-bias literature are consistent with this.

Proposition 3 (accuracy paradox). Write checking as a function of the error rate, $c(e)$, rising as errors become more frequent, and let ε be the elasticity of checking with respect to the error rate, that is, the percentage change in checking for a one per cent change in error. Differentiating $H = e(1 - d \cdot c(e))S$ shows that reducing the error rate *increases* expected undetected harm whenever

$$\varepsilon > (1 - c \cdot d) / (c \cdot d)$$

The condition is easy to meet precisely where vigilance is currently high and effective. If users check and detect nine errors in ten ($c \cdot d = 0.9$), an elasticity of just 0.11 is enough for an improvement in accuracy to make things worse at the margin. If they catch only half ($c \cdot d = 0.5$), the elasticity must exceed 1. The paradox is thus sharpest in the settings we

would least expect: those in which careful professionals currently catch most errors and a better system persuades them to stop. This is the formal shadow of Dell'Acqua's recruiters, and of Bainbridge's irony.

5.5 Waking trust and sleeping trust

What kind of state is a low c ? It is tempting to call it trust and to make trust the problem. That would be a mistake. No one can verify everything, and trust is not optional. The line that matters runs not between trust and distrust but between two kinds of trust.

Waking trust is reliance whose monitoring dispositions remain live. It does not check everything, but it would notice and respond to signs that something is wrong: an answer that does not fit, stakes that call for confirmation, a source straying beyond its competence. It has counterfactual sensitivity: had things been otherwise, the truster would have noticed. *Sleeping trust* is reliance whose monitoring dispositions have been disengaged. It continues by habit; signs of error do not register because nothing is listening for them.

Annette Baier's account of trust as "accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one" already makes room for the distinction, since she notes that "there is such a thing as unconscious trust" (Baier, 1986, p. 235). Sleeping trust is the kind of unconscious trust produced not by long familiarity with a tested good will but by the presentational features of a source that has no will at all. Table 2 sets out the contrast.

Table 2. Waking trust and sleeping trust

Dimension	Waking trust	Sleeping trust
Monitoring dispositions	Live: would notice and respond to signs of error	Disengaged: signs of error do not register
Ground of reliance	Track record, institutions, reasons that can be inspected	Presentational cues: fluency, assurance, completeness
Relation to stakes	Rises to confirmation when stakes rise	Indifferent to stakes; the same habit everywhere
Reasons	Accessible; can be asked for and assessed	Not sought; the verdict suffices
Phenomenology	Confidence held with a residual readiness to doubt	Relief; closure; the question feels finished
Kant	Reading a book that informs my understanding	"A book that understands for me"

Dimension	Waking trust	Sleeping trust
Ibn Rushd (§9)	Accept and thank for what agrees with the truth; alert to what does not	Accept whatever is said

The two kinds of trust may involve the same amount of reliance on any given occasion. The difference is not how much one relies but whether the capacity to stop relying is still awake. Relief is the moment trust falls asleep.

5.6 The vigilance-preservation criterion

We can now state the normative proposal.

Vigilance-preservation criterion. An AI system should be evaluated not only by its accuracy but by its effect on the effective vigilance of its users, weighted by stakes. A system fails the criterion to the extent that it lowers users' probability of checking and detecting its errors, in decisions where the stakes warrant checking, below the level that the system's error profile requires.

Three corollaries follow. First, evaluation should report the joint human–system rate of undetected error, $\Sigma e(1 - c \cdot d)S$ across decision types, not only model accuracy. Second, a system should *direct* vigilance to where its errors are likely; that is the real point of calibration, which is valuable because it raises c on the outputs that most need it. Third, the criterion is proportional. It does not demand maximal friction, which would waste attention on low-stakes questions; it demands that vigilance be preserved where it matters.

6 Trusteeship: why responsibility cannot be borrowed

6.1 The trusteeship paradigm

Why is lost vigilance a *moral* failure rather than a misfortune, and why can responsibility not simply pass to the system or its maker? For this we turn to a resource the philosophy of technology has largely overlooked: the trusteeship paradigm (*al-i'timāniyya*) of the Moroccan philosopher Taha Abdurrahman (Ṭāhā 'Abd al-Raḥmān).

Taha's ethics, as reconstructed by Hashas (2020) and examined at length by Hallaq (2019), rests on three principles: *shahāda* (testimony or witnessing), *amāna* (trust) and *tazkiya* (self-purification). *Amāna* takes its name from the Qur'ānic verse in which the trust was offered to the heavens, the earth and the mountains, which declined to bear it, while the human being accepted it (Q 33:72). On Taha's reading, to have accepted the trust is to be free and accountable at once. To fail in it is either *khiyāna*, betrayal of what was entrusted, or *ḥiyāza*, possession of what is not one's own as though it were (Hashas, 2020). Taha insists that there is no human being without ethics (ʿAbd al-Raḥmān, 2000, as

discussed in Hashas, 2020), and he distinguishes a jurisprudence of command (*fiqh i'timārī*), which treats obligation as external rule-following, from a jurisprudence of trusteeship (*fiqh i'timānī*), in which obligation is borne from within as a trust (Hashas, 2020). He has also turned this framework explicitly towards the ethical challenges of the media and communication revolution, in the second volume of *Dīn al-ḥayā'* (Abd al-Raḥmān, 2017; see Hashas, 2020).

6.2 The epistemic trust cannot be delegated

The trusteeship paradigm offers a precise way of saying what the vigilance-preservation criterion protects. Among the things a human being bears as a trust is responsibility for what he or she assents to and acts upon. Call this the *epistemic amāna*. A trust in this sense is not a task that can be subcontracted. Tasks can be delegated, and much of good epistemic life consists in delegating them well. But the standing of the trustee, the one who answers for the assent, is not a task. It is a relation that constitutes the person as accountable, and it does not move when the work moves.

When we adopt a machine's confident answer without exercising judgement over it, I have not transferred my responsibility to the machine. I cannot. I have left the watch unkept. In Taha's vocabulary, which we are here extending beyond its original use, borrowed certainty has the structure of *ḥiyāza*: I hold a certainty as though it were my own when it is not. And sleeping trust has the structure of *khiyāna*, not as wilful betrayal but as a trust left unguarded. The relief that accompanies a fluent answer is the feeling of having put down a burden. The trusteeship paradigm says that this particular burden cannot be put down. It can only be neglected.

6.3 The responsibility gap, seen from the other side

Matthias (2004) argued that learning machines create a *responsibility gap*, in which neither manufacturer nor operator can fairly be held responsible for behaviour neither could predict or control. That debate concerns who answers for what the machine does. The trusteeship argument concerns something else: who answers for what *I* believe and do on the basis of what the machine says. Even if the gap for the machine's outputs were unbridgeable, my trusteeship over my own assent would be untouched. The machine can be a source but, on Taha's framework, not a trustee, because a trust is accepted by a being that can answer for it and betray it; the Qur'ānic image is of entities that could not take it on. This is a claim not about machine consciousness but about the structure of accountability, which attaches to the one who assents.

6.4 Ibn Rushd: demonstration, rhetoric and the deposit

A second Arabic source sharpens the epistemic side of the argument. In the *Decisive Treatise*, Ibn Rushd (Averroes) defines consideration as "nothing more than inferring and drawing out the unknown from the known" (Ibn Rushd, 2001, §3) and observes that "people's natures vary in excellence with respect to assent. Thus, some assent by means of demonstration; some assent by means of dialectical statements ...; and some assent by means of rhetorical statements" (§11). Certainty in the strict sense belongs to demonstration: the jurist "has only a syllogism based on supposition, whereas the one who is cognizant has a syllogism based on certainty" (§14).

A language model's output is, structurally, rhetorical: it produces assent through persuasive presentation, and its user rarely has access to demonstration of its claims. Rhetorical assent is not wrong as such; Ibn Rushd thought it the proper route for most people on most matters. The danger is a confusion of kinds: rhetorical persuasion experienced as demonstrative certainty. Borrowed certainty is exactly that confusion, a feeling proper to demonstration produced by means proper to rhetoric.

Ibn Rushd also offers a model of waking trust. Discussing the books of the Ancients, he writes that "we will accept, rejoice in, and thank them for whatever agrees with the truth; and we will alert to, warn against, and excuse them for whatever does not agree with the truth" (§9). This is reliance with judgement retained: generous, grateful and alert. His treatment of error has the same structure: the adept who errs is excused, like a skilled physician, but "one who is not adept in that concern is not excused" (§25). Responsibility follows the exercise of competence, as it does in Sosa's account of aptness.

Finally, Ibn Rushd reaches for the same verse as Taha. "Sound interpretation", he writes, "is the deposit mankind was charged with holding and held, whereas all existing things shirked it" (§51, citing Q 33:72). For Ibn Rushd, the trust humanity accepted is the labour of interpretation itself, the work of judgement.

6.5 An honest note on the conversation

We should not overstate the harmony between these two thinkers. Taha is a sharp critic of Ibn Rushd, whom he regards as an imitator of Aristotle rather than an original philosopher (Hashas, 2020, p. 40). Ibn Rushd's scheme is also openly hierarchical, reserving demonstration for a few. The convergence we claim is limited and specific: both read the Qur'ānic trust as something borne by the human being that no other entity accepted, and both tie it to the exercise of judgement. The synthesis is ours. We also depart from Ibn Rushd's elitism in a Kantian direction. The point of vigilance-preserving design is to extend the conditions of waking trust to everyone, not to accept that the many will assent by rhetoric while the few demonstrate.

What the Arabic thread adds is a vocabulary for non-transferability. Western discussions tend to treat responsibility as a quantity allocated among designers, deployers and users; the trusteeship paradigm treats it as a relation that constitutes the agent. On that view, "who is responsible when the AI is wrong?" has, for the user's own assent, an answer that does not depend on the AI at all.

7 Objections and replies

7.1 "We already trust experts"

Objection. We rely on doctors, encyclopaedias and pilots without checking them, and Hardwig (1985) showed that this is often rational. Borrowed certainty looks like a new name for ordinary epistemic dependence.

Reply. Dependence is indeed unavoidable and often rational, which is why the target is sleeping trust, not trust. But the cases differ in three ways. First, trust in experts is embedded in institutions that keep it awake on our behalf: credentials, professional discipline, liability, second opinions, and the expert's capacity to give reasons and to say "I don't know". Second, an expert's confidence is linked, imperfectly, to a track record in a domain. An LLM's presentational confidence is substantially decoupled from its reliability (Section 4.4), and the jagged frontier (Dell'Acqua et al., 2026) means that reliability on one task says little about the next. Third, a general-purpose system is consulted on everything in one voice, so a single relief profile governs dependence across domains in which no human expert would claim competence. On Goldberg's (2010) social reliabilism, whether reliance is good depends on the whole system of sources and safeguards, and here most of the safeguards that make expert dependence reasonable are missing.

7.2 "Calibration solves it"

Objection. Make systems calibrated, so that expressed confidence matches the probability of being right, and users will check when they should.

Reply. Calibration is necessary but not sufficient. First, calibration belongs to the machine and vigilance to the human, and the evidence shows a gap between them: users discriminate correct from incorrect answers much worse than the model's own confidence would permit (Steyvers et al., 2025), and uncertainty expressions reduce but do not eliminate over-reliance (Kim et al., 2024). Second, the pipelines that shape model behaviour appear to penalise hedging (Zhou et al., 2024), so calibration has to be fought for against the grain of user preference. Third, and most fundamentally, a perfectly calibrated system that is usually right will usually, and correctly, say that it is confident, which is a ra-

tional invitation not to check. Calibration raises checking on outputs flagged as uncertain. It does nothing for errors that arrive with high confidence, and in a well-calibrated system that is usually right, some always will.

7.3 "Friction harms usability"

Objection. The proposal makes AI worse to use. The designs that reduced over-reliance most were the least liked (Bućinca et al., 2021), and users will abandon them for frictionless rivals.

Reply. This is the strongest objection. First, the criterion demands proportional friction; for most low-stakes queries the right amount is close to none. Second, the objection takes user satisfaction as the measure of a good system, but satisfaction is precisely what relief produces, and a metric that rewards relief will select for borrowed certainty. Third, friction can be targeted at the moments that matter (irreversible actions, high-stakes decisions, outputs the system rates as uncertain) and can take forms that lower the cost of checking, the mechanism Vasconcelos et al. (2023) identified for explanations that work.

7.4 "Offloading is simply how cognition works"

Objection. Humans have always offloaded cognition onto notes, calculators and maps, a normal and often adaptive strategy (Risko & Gilbert, 2016).

Reply. Offloading the *storage* or *computation* of information is not offloading the *judgement* of whether to accept it. A calculator does not persuade me, and its reliability in its domain is near total and well understood. The criterion is indifferent to offloading where errors are rare, cheap or easily detected. It bites where errors are non-trivial, stakes are high and presentation disables detection together, a combination characteristic of fluent general-purpose assistants and not of most traditional tools.

7.5 "Vigilance cannot be measured, and the model is too simple"

Objection. Accuracy can be benchmarked; vigilance is vague. And the model ignores human error and the costs of checking.

Reply. Vigilance here is behavioural and measurable: verification rates, verification latency and, above all, detection of errors deliberately placed, with informed consent, in evaluation settings. Several studies in Section 4 already measure versions of these. The model is simple by design. A fuller one would add checking costs, human error on correct outputs and correlations between confidence and error; none of these removes the point that vigilance multiplies every error.

8 Design and policy implications

8.1 Design: bringing back deliberate friction

The design implication of the argument can be stated in a sentence: bring back deliberate friction, in proportion to the stakes. Table 3 summarises six principles, the term in the model each targets, and the evidence that supports or complicates it.

Table 3. Vigilance-preserving design principles

Principle	Model term targeted	Supporting or complicating evidence	Main risk
D1 Honest uncertainty (unsolicited, calibrated, first-person where apt)	Raises c on uncertain outputs	Kim et al. (2024); Zhou et al. (2024); Steyvers et al. (2025)	Reduces but does not eliminate over-reliance
D2 Reasons, not just verdicts (checkable sources, cruxes)	Lowers cost of checking; raises d	Vasconcelos et al. (2023); Steyvers et al. (2025) on length inflating confidence	Ornamental explanations can increase borrowed certainty
D3 Confirmation of what matters (stakes-weighted)	Raises c where S is high	Buçinca et al. (2021)	Disliked by users if applied uniformly
D4 Targeted vigilance	Raises c on likely errors	Goddard et al. (2012): updated confidence levels as mitigator	Depends on good calibration
D5 Measured vigilance	Makes c-d visible in evaluation	Measures used in Section 4 studies	Probes only with consent, in testing
D6 Do not optimise for relief	Protects honest confidence cues	Zhou et al. (2024): bias against uncertainty in preference data	Commercial pressure towards satisfaction metrics

Three points deserve emphasis. Because a plain statement is read as a confident one (Zhou et al., 2024), the absence of a hedge is itself a signal, and a system should not send it falsely. Because longer explanations inflate confidence without improving discrimination (Steyvers et al., 2025), reasons must be checkable rather than ornamental: sources that can be opened, the crux on which the answer turns, and what would change it. And *measured vigilance* means consented error-detection probes in testing, never deceiving users in deployment.

8.2 Policy

The European Union's AI Act already recognises part of the problem. Article 14(4)(b) requires that high-risk AI systems be provided in such a way that the persons assigned to human oversight are enabled "to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)" (European Parliament & Council of the European Union, 2024, Art. 14(4)(b)). But awareness is not vigilance: instructions and training do not reliably prevent over-reliance (Parasuraman & Manzey, 2010), and the user who most needs the warning is the one whose relief threshold has already been crossed. Three extensions follow. Oversight obligations should be judged by evidence that vigilance is preserved, such as measured detection rates, not by warnings alone. Organisations procuring AI for high-stakes uses could require a *vigilance impact assessment* alongside the usual accuracy figures. And the logic of Article 14 should inform the use of general-purpose assistants in high-stakes settings that fall outside the Act's high-risk categories.

Goddard et al.'s (2012) identification of user accountability as a candidate mitigator connects policy to trusteeship: institutions can make the epistemic *amāna* visible by making clear, in workflow and law, that the user who acts on an AI output answers for that act. Kant asked for courage; a modern curriculum might ask for a specific, teachable form of it: the habit of asking, before acting on a fluent answer, what it rests on and what it would cost if it were wrong.

9 Proposed empirical agenda

The argument yields testable hypotheses. None has been tested in the form given here.

H1 (Relief). At equal accuracy, fluent unhedged outputs reduce felt uncertainty more than hedged ones, and this reduction mediates reduced verification. Design: manipulate hedging and fluency between subjects; measure felt uncertainty before and after the answer and behavioural verification; test mediation.

H2 (Relief threshold). The probability of verification falls non-linearly as confidence cues strengthen, with a threshold whose location shifts upwards with stakes and downwards with time pressure. Design: graded confidence cues within subjects under varied stakes and time pressure; fit a psychometric function to verification.

H3 (Accuracy–vigilance trade-off). Raising accuracy from about 90% to about 99% lowers checking enough that joint undetected error falls less than proportionally and, among experienced users, may rise. Design: extend Dell'Acqua's (2022) paradigm to an LLM assistant with seeded errors of known frequency; estimate *c*, *d*, joint undetected error and the elasticity in Proposition 3.

H4 (Sleeping trust). After a long run of correct outputs, users' detection of the first error is lower than detection of an identical error presented early in the interaction, and this effect is larger for fluent outputs. Design: manipulate the position of the first seeded error in an extended session.

H5 (Proportional friction). Friction targeted at high-stakes and low-confidence outputs preserves error detection there at lower cost in satisfaction and time than uniform friction. Design: compare no, uniform and targeted friction on detection, time and satisfaction.

H6 (Trusteeship framing). Framing that makes the user's responsibility explicit, including framing in terms of *amāna* for participants for whom that concept is meaningful, increases verification on high-stakes items without reducing appropriate reliance on low-stakes ones. Design: compare neutral, accountability and trusteeship framings in Arabic- and English-speaking samples.

In all six, the key outcome is neither accuracy nor satisfaction but the stakes-weighted joint rate of undetected error.

10 Conclusion

The common picture of AI risk is a picture of a machine that is wrong. This paper has argued for adding a picture of a person who stops looking. The confident, fluent answer relieves the discomfort of not knowing, and relief is the moment at which vigilance is lost. The result is borrowed certainty, the state of being certain without having achieved knowledge, which is real as a feeling, brittle as a belief and invisible to the person who holds it.

Certainty is a state and knowledge an achievement, so a machine that produces certainty does not thereby produce knowledge. Trust is necessary, but the line that matters runs between waking trust and sleeping trust. Because harm depends on unchecked error, a highly accurate system that extinguishes checking can be more dangerous than a less accurate one that keeps users alert, under conditions that can be stated precisely. And because responsibility for one's own assent is a trust that cannot be handed over, lost vigilance is not only a risk but a failure of that trust.

The design conclusion is modest: honest uncertainty, reasons and not just verdicts, and confirmation of what matters. The ethical conclusion is less so. A system that speaks to human beings cannot be judged by its accuracy alone, but also by what it leaves awake in the person who listens.

We may lend our questions to machines, but we must never lend them our judgement. Certainty can be borrowed. Responsibility cannot.

References

- ‘Abd al-Raḥmān, Ṭ. [Taha Abdurrahman]. (2000). *Su’āl al-akhlāq* [The question of ethics]. al-Markaz al-Thaqāfī al-‘Arabī.
- ‘Abd al-Raḥmān, Ṭ. [Taha Abdurrahman]. (2017). *Dīn al-ḥayā’* [The religion of modesty] (Vols. 1–3). al-Mu’assasa al-‘Arabiyya li-l-Fikr wa-l-Ibdā’.
- Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3), 219–235. <https://doi.org/10.1177/1088868309341564>
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2), 231–260. <https://doi.org/10.1086/292745>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- 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), Article 188, 1–21. <https://doi.org/10.1145/3449287>
- Dell'Acqua, F. (2022). *Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters* [Unpublished working paper]. <https://www.fabriziodellacqua.com/>
- Dell'Acqua, F., McFowland, E., III, Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2), 403–423. <https://doi.org/10.1287/orsc.2025.21838>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Fromm, E. (1941). *Escape from freedom*. Farrar & Rinehart.
- Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), Article 6. <https://doi.org/10.3390/soc15010006> (Correction published 2025, *Societies*, 15(9), Article 252, <https://doi.org/10.3390/soc15090252>)
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>

Goldberg, S. C. (2010). *Relying on others: An essay in epistemology*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199593248.001.0001>

Hallaq, W. B. (2019). *Reforming modernity: Ethics and the new human in the philosophy of Abdurrahman Taha*. Columbia University Press. <https://doi.org/10.7312/hall19388>

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7), 335–349. <https://doi.org/10.2307/2026523>

Hashas, M. (2020). The trusteeship paradigm: The formation and reception of a philosophy. In M. Hashas & M. al-Khatib (Eds.), *Islamic ethics and the trusteeship paradigm: Taha Abderrahmane's philosophy in comparative perspectives* (pp. 37–61). Brill. https://doi.org/10.1163/9789004438354_004

Ibn Rushd [Averroes]. (2001). *The book of the decisive treatise determining the connection between the law and wisdom, and Epistle dedicatory* (C. E. Butterworth, Trans.). Brigham Young University Press. (Original work written c. 1179)

Kant, I. (1996). An answer to the question: What is enlightenment? In M. J. Gregor (Ed. & Trans.), *Practical philosophy* (pp. 11–22). Cambridge University Press. <https://doi.org/10.1017/CBO9780511813306.005> (Original work published 1784; cited by Akademie Ausgabe [AA] volume and page)

Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Wortman Vaughan, J. (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* (pp. 822–835). ACM. <https://doi.org/10.1145/3630106.3658941>

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, 103(2), 263–283. <https://doi.org/10.1037/0033-295X.103.2.263>

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121, pp. 1–22). ACM. <https://doi.org/10.1145/3706598.3713778>

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. <https://doi.org/10.1007/s10676-004-3422-1>

- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49–58. <https://doi.org/10.1038/s41586-024-07146-0>
- Nguyen, C. T. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplement*, 89, 227–255. <https://doi.org/10.1017/S1358246121000035>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Popper, K. R. (2002). *Conjectures and refutations: The growth of scientific knowledge*. Routledge. (Original work published 1963)
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006. <https://doi.org/10.1006/ijhc.1999.0252>
- Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. 1). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199297023.001.0001>
- Sosa, E. (2011). *Knowing full well*. Princeton University Press. <https://doi.org/10.23943/princeton/9780691143972.001.0001>
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>
- Steyvers, M., Tejada, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231. <https://doi.org/10.1038/s42256-024-00976-7>
- Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), 1–38. <https://doi.org/10.1145/3579605>
- Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3623–3643). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.198>