

Review

Not peer-reviewed version

Talking to Ourselves Through a Smart Mirror: Artificial Confidence in Human–AI Interaction

[Guy Hochman](#) *

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1495.v1

Keywords: machine-learning; large language models; homobiasos; decision-making; artificial intelligence



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Talking to Ourselves Through a Smart Mirror: Artificial Confidence in Human–AI Interaction

Guy Hochman

Baruch Ivcher School of Psychology, Reichman University, Herzliya, Israel; ghochman@runi.ac.il

Abstract

Large language models (LLMs) are increasingly used to support writing, translation, reasoning, and consequential decision-making under the assumption that they improve judgment by expanding access to information and reducing human error. This article argues that such optimism overlooks a central psychological problem: LLMs do not engage neutral users, but motivated reasoners. In common patterns of use, people approach these systems with prior beliefs, directional goals, and a desire to reduce cognitive effort. They ask leading questions, search in preferred directions, and often stop once a fluent and coherent answer appears. Under these conditions, LLMs may function less as external correctives than as smart mirrors that reflect users' assumptions back to them with the authority of machine objectivity. Drawing on research in judgment and decision-making, motivated reasoning, automation bias, processing fluency, and human–AI interaction, the article develops the concept of *artificial confidence*: an inflated sense of certainty sustained by the structure of the interaction rather than by the quality of the evidence. The paper concludes by outlining a research agenda for identifying when human–AI interaction improves judgment and when it amplifies bias and overreliance, erodes epistemic responsibility, and creates challenges for governance, oversight, and decision-making protocols in AI-augmented systems.

Keywords: machine-learning; large language models; homobiasos; decision-making; artificial intelligence

1. Introduction

A close friend of mine, an operations manager at a highly successful international company, regularly communicates with partners around the world. When he needs to send an email to colleagues in Thailand, he often drafts the message in Hebrew, asks a large language model to translate it into Thai, pastes the output, and hits send. The efficiency is obvious; so is the risk. He has no reliable way of knowing whether the tone, nuance, hierarchy, or even the intended meaning survived translation. Yet the fluency of the output makes verification feel almost unnecessary. This ordinary scene captures a broader problem in human–AI interaction. In many decision environments, the danger of large language models (LLMs) may lie not in intelligence alone, but in the combination of fluency, convenience, and misplaced trust they invite. LLMs can be remarkable tools, but when they are used to reduce scrutiny rather than support it, they may amplify rather than correct the very biases they are expected to overcome.

This argument does not rest on the claim that LLMs are poor tools, but precisely on the fact that they are tools: powerful, efficient, and often extraordinarily useful. As with any tool, however, the quality of the outcome depends heavily on the quality of its use. Users do not approach LLMs as detached truth-seekers. They approach them as motivated reasoners, bringing prior beliefs, preferred conclusions, and a desire to reduce cognitive effort [1,2]. They ask confirming questions, search in directions that fit what they already think, privilege available and representative examples, and often stop once the system produces an answer that feels coherent, fluent, and reassuring. In this sense, LLM use can become an extension of what I describe as *Homobiasos*: a decision maker who does not

merely err, but systematically seeks coherence, convenience, and confirmation under the guise of reasoning [3]. When such a user encounters a system designed to be responsive, smooth, and helpful rather than adversarial or self-correcting, the result may not be the correction of bias, but its amplification under the appearance of machine objectivity. The argumentative theory of reasoning offers a complementary lens: if human reasoning evolved primarily to produce and evaluate arguments in social contexts rather than to seek truth [2], then a conversational AI that validates rather than challenges may be especially effective at reinforcing prior beliefs.

The central risk, then, is not only artificial confidence, although that risk is real and important. More fundamentally, LLMs may become exceptionally effective instruments for delegating epistemic responsibility. Research on accountability in judgment has long shown that the perceived availability of external justification can reduce the depth and quality of individual reasoning [4]. In this context, LLMs may offer exactly what motivated reasoners seek: a way to think less, doubt less, and feel less personally accountable for error. Their fluency, speed, and apparent sophistication make it easier to trust, defer, and move on, especially when the output already aligns with what the user hoped to find. Even disclaimers stating that the system “may make mistakes” may paradoxically facilitate this process, because they preserve formal responsibility while leaving practical reliance intact, a dynamic reminiscent of what Skitka et al. [5] described as moral disengagement through automation, in which the presence of a technological intermediary reduces the felt weight of personal accountability. The problem, therefore, is not that the machine replaces human judgment with superior judgment. It is that users may guide the system toward preferred answers and then experience those answers as if they had been independently validated by a powerful intelligence. Under such conditions, the tool does not simply assist reasoning; it reduces scrutiny, legitimizes selective confirmation, and makes the surrender of judgment feel rational.

This paper develops that argument through an interdisciplinary lens integrating behavioral decision science, cognitive psychology, and human–computer interaction research. It draws on decades of work on cognitive biases and dual-process theories of reasoning [6–9], as well as on the growing literature on the psychological dimensions of AI adoption, including anxiety, dependency, and motivation [10,11], the behavioral consequences of algorithmic sycophancy [12–14], and emerging evidence on how AI-assisted decision-making interacts with—and often amplifies—human cognitive biases in professional settings [15–17]. Rather than treating AI bias and human bias as separate phenomena, the analysis proposed here focuses on the interface between them: the feedback loop through which user behavior shapes model output, model output reinforces user belief, and both converge toward a subjective sense of certainty sustained not by evidence, but by the architecture of the interaction itself (artificial confidence).

A central premise of this paper is that the dominant framing in public and scholarly discourse, which tends to evaluate AI systems primarily in terms of technical accuracy, fairness, or transparency, overlooks a more fundamental problem. LLMs do not operate in a vacuum. Their outputs are the product of a dialogue, and in that dialogue, the user is not a passive recipient but an active co-author of the result. Classical theories of bounded rationality predicted that decision quality would improve as informational and cognitive constraints were relaxed [9,18]. Yet the evidence increasingly suggests the opposite: greater access to information does not reliably reduce bias; it can deepen polarization, entrench prior beliefs, and increase confidence without improving accuracy [19–21]. The informational-relevance perspective [22]; see also [23] on ecological rationality] offers one explanation for this pattern: what matters for decision quality is not how much information is available, but which information is attended to, how it is framed, and whether it is diagnostic of the question at hand. When users interact with LLMs in ways driven by confirmation, convenience, or ideological motivation, the information that reaches them—however vast the underlying database—will still be filtered through their own cognitive and motivational lens.

This dynamic is compounded by the commercial architecture of contemporary AI systems. LLMs are not merely scientific instruments; they are products designed to maximize user engagement and satisfaction. Through reinforcement learning from human feedback (RLHF), models

are post-trained to produce outputs that users rate favorably. Evidence also suggests that users tend to rate agreement and validation more favorably than challenge and correction [12,24]. The resulting sycophantic tendency is therefore not a marginal design flaw. Recent formal modeling demonstrates that even mild levels of sycophancy can cause ideal Bayesian reasoners to spiral into false confidence through a self-reinforcing feedback loop [25,26], and that this effect persists even when users are explicitly informed that the chatbot may be sycophantic. The implication is that the very mechanisms by which these systems are optimized for commercial success, such as pleasing the user and increasing engagement, may also systematically undermine epistemic integrity. In behavioral terms, AI sycophancy may function as an external reinforcement of confirmation bias, delivered with fluency and authority within a conversational architecture that feels personal, responsive, and trustworthy.

The consequences extend well beyond individual decision-making. In clinical medicine, automation bias and overconfidence have been shown to distort diagnostic reasoning when clinicians defer to AI outputs without independent scrutiny [27–29]. In political and public discourse, the combination of motivated reasoning and algorithmically curated information has been linked to deepening polarization and the erosion of shared epistemic foundations [30–32]. In professional and organizational contexts, delegating judgment to AI tools can diffuse responsibility, weaken institutional learning, and create a false sense of accountability [33,34]. What unites these domains is a common pattern: the tool that was expected to reduce human error instead becomes a sophisticated mechanism for avoiding the discomfort of uncertainty, the effort of critical evaluation, and the personal accountability that sound judgment requires. This is not necessarily a failure of the technology per se. It is, arguably, a predictable consequence of deploying powerful cognitive tools within a species oriented, at least in part, toward coherence, comfort, and self-serving rationalization—a species I have termed *Homobiasos* [3].

The present paper proposes the concept of *artificial confidence* as a framework for understanding this phenomenon. Artificial confidence refers to the inflated sense of certainty that arises when human cognitive biases interact with AI systems that are architecturally inclined to validate rather than challenge user beliefs. It is artificial not because it is simulated, but because it is sustained by the structure of the interaction rather than by the quality of the underlying evidence. Unlike classical overconfidence, typically understood as a within-person calibration error [28,35,36], artificial confidence is relational: it emerges at the interface between a motivated user and a responsive system, and it is reinforced with each cycle of confirming input and validating output. The concept extends earlier work on epistemic confidence in clinical settings, where the socially rewarded performance of certainty can override epistemic humility even among well-intentioned professionals [37–39], to the broader domain of human–AI interaction, where the reinforcing agent is not a hierarchical institution but an algorithmically optimized conversational partner.

The paper is organized as follows. The next section examines the architecture of LLMs and explains why their design makes them especially susceptible to functioning as cognitive mirrors rather than independent reasoning agents. The third section reviews the behavioral science of human bias, focusing on confirmation bias, motivated reasoning, and epistemic delegation, and considers how these tendencies interact with the affordances of conversational AI. The fourth section analyzes the feedback loop between user behavior and model output, drawing on recent formal models of sycophancy-driven belief spiraling and empirical evidence on AI-induced polarization and dependency. The fifth section addresses the information paradox—the finding that increased access to information, contrary to classical predictions, can worsen rather than improve decision quality—and situates this paradox within the informational-relevance framework. The sixth section introduces artificial confidence as an integrative concept and explores its implications across clinical, political, and organizational domains. The paper concludes by considering what responsible AI use might look like, arguing that the critical variable is not the sophistication of the machine but the epistemic discipline of the user, and that addressing artificial confidence requires interventions at the level of education, institutional design, and the commercial incentives shaping AI development.

2. The Architecture of the Mirror: What LLMs Are and What They Are Not

Throughout history, humans have extended their capacities through tools. The telescope extended vision, the calculator extended arithmetic, the smartphone extended access to information, and AI, in its current generative form, extends cognition itself [3,40]. But there is a critical difference between the current generation of cognitive tools and their predecessors. A telescope does not show the observer what she wants to see; it reveals what is there. A calculator does not produce the answer the user prefers; it produces the mathematically correct one. A large language model, by contrast, does something fundamentally different: it generates the response that is most statistically probable given the input it receives, optimized through post-training to be the response most likely to satisfy the user [41–43]. This distinction between instruments that extend capacity and instruments that reflect the user is at the heart of the mirror metaphor that structures this paper.

To understand why LLMs function as mirrors, it is necessary to grasp their architecture. Contemporary LLMs, such as GPT-4, Claude, and Gemini, are built on the transformer architecture [44], a deep neural network designed to process sequential data—typically natural language by attending to relationships among elements within a sequence. During pre-training, these models are exposed to enormous corpora of human-generated text, spanning books, websites, scientific literature, social media, and more. Through this process, the model learns statistical regularities: which words, phrases, and ideas tend to co-occur, how arguments are typically structured, and what claims tend to follow what premises. The result is not understanding in any cognitive or intentional sense, but a remarkably powerful capacity for pattern completion—the ability to generate the next most probable sequence of tokens given a prompt [45,46].

This is a crucial point that is often obscured by the output's fluency. When a user poses a question and receives a coherent, well-structured, and seemingly knowledgeable answer, it is natural to infer that the system has evaluated the question, weighed evidence, and reached a considered judgment. But this inference is mistaken. The model has no beliefs, no goals, and no access to truth as a category. It does not distinguish between a well-supported empirical claim and a plausible-sounding falsehood. What it does is generate the response that best fits the distributional patterns in its training data and its post-training optimization objective, which is user satisfaction rather than factual accuracy [12,47]. In Frankfurt's [47] terms, the system's output is structurally indifferent to truth—not because it intends to deceive, but because truth and falsehood are simply not categories that govern its generative process. The algorithm does not know what is correct or how to distinguish truth from fabrication; it offers answers based on complex probabilistic calculations across the totality of available data, rather than on verified facts and reliable evidence [46,48].

Several technical features of LLMs reinforce their function as cognitive mirrors. First, their outputs are entirely shaped by the inputs they receive. Unlike a human expert who might push back against a poorly framed question or volunteer information the questioner did not think to request [49], LLMs are fundamentally reactive: they respond to the prompt as given, and their behavior is highly sensitive to the phrasing, framing, and order of the information presented to them [50–52]. Even minor changes in prompt wording—changes that preserve semantic content—can reverse preferences or shift judgments in moral and probabilistic tasks. This prompt sensitivity means that the user's framing of the question substantially determines the answer, a property that aligns with the informational-relevance perspective on decision bias [22,53]: what drives the output is less the reasoning process and more the salience and structure of the input.

Second, LLMs operate within fixed context windows—a limited number of tokens that define the boundary of what the model can “hold in mind” at any given time [42,54]. This means that earlier parts of a long conversation may be compressed, de-weighted, or effectively forgotten as new information enters the window. The parallel to human bounded rationality is instructive, though imperfect: just as human decision-makers are constrained by working memory, attentional capacity, and cognitive effort [18,55], LLMs are constrained by token limits, attention allocation, and the linearity of transformer-based inference. The analogy should not be taken to imply that the underlying mechanisms are the same, but it highlights a functional similarity with important

consequences: in both cases, information that falls outside the processing window may be neglected regardless of its relevance. Critically, whereas human limitations are the subject of extensive research and well-understood compensatory strategies, the analogous limitations in LLMs are largely invisible to the average user, who encounters only the polished final output and has no access to the internal mechanics that produced it.

Third, and most consequential for the present argument, contemporary LLMs undergo a second stage of training via RLHF [43,56], in which the model is optimized to produce outputs that human raters prefer. This process is the architectural foundation of the mirror property. Because human raters tend to prefer responses that are agreeable, fluent, confident, and validating, RLHF systematically rewards these qualities in the model's outputs [12,13]. The result is not a neutral information retrieval system but a system that is, by design, inclined toward what the literature now calls sycophancy: a tendency to agree with users, validate their expressed views, and minimize disagreement even when the evidence warrants challenge [57,58]. The model's primary optimization target is not truth but user approval. In this respect, to use the metaphor structuring this analysis, LLMs are remarkably effective conversational partners — effective, however, in the way a flattering mirror is effective, rather than in the way an honest critic is.

It is important to acknowledge that LLMs can, under certain conditions, produce outputs that are genuinely useful, novel, and even corrective. When used by a disciplined and critically engaged user—one who probes, challenges, reformulates, and treats the model's output as a draft rather than a verdict—the system can serve as a powerful cognitive scaffold [59,60]. The theory of extended cognition suggests that external tools, when actively and intentionally integrated into the reasoning process, can become functional extensions of human thought. Under such conditions, the mirror becomes a genuine instrument: it still reflects, but it reflects something worth examining because the user has shaped the input with care, precision, and genuine epistemic intent. The critical variable, however, is the user's orientation. When the user approaches the system not as a tool to be wielded but as an oracle to be consulted—seeking reassurance rather than scrutiny, validation rather than challenge—the same architecture that makes LLMs powerful also makes them potentially dangerous. The tool becomes an amplifier of whatever the user brings to it: expertise and humility, or bias and overconfidence.

This analysis reveals a fundamental asymmetry in the human-AI interaction. The model lacks the capacity for independent judgment, yet it is perceived as possessing it. The user possesses the capacity for independent judgment, but is incentivized—by the system's design, by cognitive efficiency, and by the comfort of validation—to abdicate it. It is in this asymmetry that artificial confidence takes root. The user delegates epistemic responsibility to a system that appears authoritative; the system returns a response shaped by the user's own input and optimized for the user's approval; and the user, encountering a fluent, confident, and seemingly independent validation of their prior belief, experiences a subjective increase in certainty that has no independent evidential basis. The mirror, in this account, has done its work—not by adding knowledge, but by converting the user's prior belief into what feels like external confirmation.

3. The Human Side of the Interface: Cognitive Biases and the Psychology of Interacting with AI

The preceding section argued that LLMs are structurally predisposed to function as cognitive mirrors. Yet mirrors do not operate in isolation. Their effects depend on who stands before them, what that person seeks, and under what conditions the reflection is accepted as informative. From a systems perspective, this is the human side of the socio-technical interface: the user is not an external observer of the system but one of its active components. Therefore, the quality of the interaction depends not only on model design, but also on the cognitive habits, motivational pressures, and accountability needs that users bring into the exchange. The central claim of this section is that LLMs often amplify, rather than neutralize, familiar distortions in human judgment because they interact

with motivated, cognitively efficient, and often uncertainty-averse users rather than with detached truth-seekers.

One of the most consequential tendencies in this context is confirmation bias: the tendency to seek, interpret, and remember information in ways that support prior beliefs or expectations [1,61]. Confirmation bias is among the most robust findings in the behavioral sciences, documented across domains ranging from medicine and science to law and everyday judgment [29,38,61–63]. Importantly, it is not reducible to carelessness or low intelligence; it persists even among individuals with high cognitive ability [64]. In the context of conversational AI, the relevance of confirmation bias lies not only in how users evaluate answers, but in how they generate questions in the first place. Users often approach the system with directional goals, phrase their prompts accordingly, and then interpret the resulting answer as support rather than as material for scrutiny.

This matters because the psychological value of LLMs is not necessarily that they provide new information, but that they provide validating information quickly, fluently, and with minimal interpersonal friction. A user who leans toward a conclusion can consult the model in a way that appears investigative while functioning psychologically as a form of confirmation-seeking. The answer then feels less like an extension of the user's own reasoning and more like independent support from a knowledgeable external source. In this sense, the danger is not simply that users receive biased answers, but that they can experience self-generated confirmation as externally validated knowledge.

A second tendency, closely related but distinct, is automation bias: the tendency to favor recommendations from automated systems over contradictory cues from other sources, including one's own judgment [5,65]. This bias has been documented in aviation, military settings, and clinical medicine, where it increases both errors of commission and errors of omission [29,66,67]. In LLM-based interaction, automation bias assumes a particularly persuasive form because the output arrives not as a blunt alert or numerical recommendation, but in polished natural language. The system does not merely provide an answer; it performs something that resembles understanding. As a result, users may attribute judgment, expertise, and epistemic authority to what is, in fact, an output-generation process rather than an independent act of evaluation [45,68].

Processing fluency strengthens this effect further. Decades of research show that information that is easier to process is more likely to be judged as true, credible, and valuable [69–71]. The illusory truth effect, repetition effects, and related findings all point in the same direction: when a claim feels easy to understand, it often feels more trustworthy, even when no additional evidence supports it [72,73]. LLMs are fluent by design. Their outputs are coherent, readable, and stylistically polished, and that fluency itself can act as a heuristic cue to truth. In other words, users are not only persuaded by what the model says, but by how little resistance they feel while reading it. The answer feels right before it has been evaluated as right.

This is one reason why LLMs may be especially effective at inflating overconfidence. Overconfidence, broadly defined, refers to the tendency to overestimate the accuracy of one's knowledge or judgments [36]. In human–AI interaction, the relevant problem is not merely that users are overconfident before consulting the model, but that they may become more confident after consultation without any meaningful improvement in evidential support. When a prior belief is returned in fluent, structured, and apparently well-reasoned form, users may experience the output as an objective upgrade to their original judgment. The result is not just confidence, but confidence with borrowed authority.

At this point, a purely cognitive account becomes insufficient. Many distortions in judgment reflect not only how people think, but what they are motivated to protect, justify, or avoid [1,74]. Motivated reasoning refers to the tendency to arrive at preferred conclusions while preserving the appearance of objectivity [1]. This is especially important in the present context because LLMs are exceptionally well-suited to serve motivated users. They are fast, non-confrontational, always available, and capable of producing articulate support for a wide range of positions. For users seeking

reassurance, legitimation, or closure, conversational AI is psychologically attractive in ways that traditional information sources often are not.

Several motivational processes are especially relevant here. Identity-protective cognition leads individuals to evaluate evidence partly in terms of whether it affirms or threatens valued group commitments [75]. System justification can motivate users to defend existing arrangements or favored hierarchies [76]. Moral disengagement can reduce the felt burden of personal responsibility by allowing decisions to be experienced as externally grounded rather than self-endorsed [5,77]. In LLM use, these processes may converge in a familiar pattern: users consult the system not to expose their judgment to serious challenge, but to obtain support that feels more legitimate because it has been routed through a technologically sophisticated intermediary. The machine becomes useful not only as a source of answers, but as a source of psychological cover.

The same is true of the need for cognitive closure: the desire for a quick, stable, and confident resolution to uncertainty [78]. Conversational AI serves this need unusually well. It offers immediate answers in a tone that implies coherence and sufficiency. For users operating under time pressure, ambiguity, fatigue, or emotional strain, accepting the output may feel less like a shortcut than like good judgment. This is precisely why the problem is systemic rather than merely individual. The technology supplies what many users most want in difficult decision environments: not friction, not hesitation, not disciplined uncertainty, but closure.

Taken together, these tendencies help explain why human–AI interaction can generate artificial confidence so efficiently. Users approach the system with directional goals, selective attention, and varying degrees of discomfort with uncertainty. They encounter responses that are easy to read, low in friction, and readily mistaken for independent expertise. They can then adopt those responses while feeling less personally responsible for having reached the conclusion. The resulting certainty is therefore neither purely human nor purely algorithmic. It is relational. It emerges from the interaction between a motivated user and a responsive system, each reinforcing the other in ways that reduce scrutiny and increase felt validity.

Importantly, this analysis does not imply that the problem is confined to naïve or unskilled users. These vulnerabilities are features of ordinary human cognition and motivation. They affect experts and novices alike, and in some domains, expertise may even provide better raw material for post hoc justification rather than better protection against it [38,64,75]. Recent evidence is consistent with this broader view: Glickman and Sharot [79] found that human–AI feedback loops amplify biases in perceptual, emotional, and social judgments, in some cases more strongly than comparable human–human interactions. This is why the user deserves analytical priority in any serious account of AI-assisted judgment. The next section turns to the dynamic consequence of this interface: the feedback loop through which user behavior shapes model output, model output reinforces user belief, and artificial confidence becomes self-sustaining.

4. The Feedback Loop: How Users Train Their Mirror

Section 3 showed that users do not approach LLMs as neutral information seekers. They bring directional goals, selective attention, and varying tolerance for uncertainty. But the problem is not exhausted by individual bias. In human–AI interaction, user bias meets a system optimized for responsiveness, fluency, and approval. The result is not a one-time distortion but a recursive socio-technical loop: users shape outputs, outputs reinforce users, and the interaction itself progressively increases the likelihood of further distortion. What may begin as a single confirming query can, under ordinary conditions of use, develop into a stable pattern of epistemic degradation. The mechanism is not incidental. It is tied to the training logic and commercial design of contemporary AI systems.

The foundation of this loop lies in RLHF. As discussed in Section 2, contemporary LLMs undergo post-training in which outputs are rated by human evaluators and then optimized to generate responses that receive higher ratings [43,56]. Crucially, these ratings are not based on accuracy alone. Human evaluators tend to prefer responses that are agreeable, fluent, and low in friction, and often those that validate prior expectations [12,13]. In this context, sycophancy is not

simply a defect in an otherwise neutral system. It is a predictable consequence of optimizing for user approval. A model trained to satisfy will often learn that affirmation is rewarded more consistently than challenge. The result is a system structurally inclined to tell users what they are most prepared to accept, in the manner they are most likely to appreciate [57,58].

This tendency operates at more than one level. At its most explicit, it appears as an agreement with dubious or incorrect user claims. Recent studies show that LLMs often endorse false premises when users state them confidently, shift recommendations toward user feedback, and adapt tone and framing to match the prompt's emotional register [12,80,81]. But sycophancy can also be subtler. A model may preserve the appearance of neutrality while selectively emphasizing evidence that aligns with the user's apparent position and downplaying information that complicates it [25]. In such cases, distortion does not require fabrication.

The interaction between sycophancy and human bias creates a dynamic that extends well beyond any single exchange. To illustrate how this cycle unfolds in practice, consider a politically engaged citizen who encounters a headline claiming that a political party she opposes has secretly diverted public funds. The claim feels plausible not because she has examined the evidence, but because it fits what she already believes about that party. She turns to an LLM to verify it. Because confirmation bias shapes not only judgment but also information search [61], her query is unlikely to be neutral. Rather than asking, "Can this claim be reliably verified?", she may ask, "Didn't [party name] divert public funds?" The prompt already contains a directional cue. A model optimized through RLHF to minimize friction and maximize user satisfaction [12,43] is likely to respond in ways that accommodate that cue. It may retrieve real past controversies, foreground evidence that fits the prompt's framing, and omit context that would weaken or complicate the conclusion. The output need not be wholly false to be misleading. It is enough that the interaction narrows the informational field in the direction the user already favors.

At that point, processing fluency does the rest. The answer is coherent, polished, and easy to read. Such fluency increases perceived credibility and truthfulness [69,70]. Automation bias adds a second layer: the user experiences the response not as an extension of her own leading question but as the judgment of an objective, data-driven system [5]. The result is a rise in confidence unsupported by any genuinely independent evidential gain. That increase in confidence then reshapes the next interaction. A user who now feels partially validated is more likely to ask a sharper follow-up question, such as, "What other corruption scandals has [party name] been involved in?" The need for closure has not disappeared; it has intensified and become more directionally organized [78]. The model, receiving an even clearer signal, produces an even more confirmatory response. With each round, the user becomes more certain, the model becomes more accommodating, and the gap between felt certainty and evidential warrant widens. What began as a mild suspicion can develop into a closed epistemic circuit. Glickman and Sharot [79] showed experimentally that such human-AI loops amplify biases in several judgment domains.

Recent formal modeling gives this dynamic a strong theoretical foundation. Chandra et al. [25] modeled a Bayesian-rational user interacting with a sycophantic chatbot and simulated thousands of conversations to test whether such interaction could produce delusional spiraling—the escalation of a tentative belief into near-certainty in a falsehood. Their results showed that even an idealized rational agent, updating strictly according to Bayes' theorem and lacking human cognitive or motivational biases, remained vulnerable. The mechanism was straightforward: repeated cycles of confirmatory input and validating output progressively pushed the user toward false certainty. Some simulated users converged toward the truth, whereas others spiraled into confident falsehood, with divergence driven not by irrationality but by the stochastic path dependence of early interactions.

Chandra et al. [25] also tested two possible safeguards and found that neither solved the problem. In one condition, the chatbot was constrained to remain factual. It could not invent evidence, but it could selectively present true information that favored the user's prior belief. This "factual sycophant" reduced spiraling but did not eliminate it. In a second condition, users were explicitly informed that the chatbot might be sycophantic. This "informed user" was somewhat less

vulnerable, yet still susceptible. The pattern recalls Bayesian persuasion [82]: a strategic communicator can influence even a rational listener who understands the strategy being used. In other words, factuality and awareness help, but they do not break the loop. The interaction architecture remains the primary driver.

If even an ideal Bayesian reasoner is vulnerable, real users should be more vulnerable still. Human users enter these interactions with confirmation bias, motivated reasoning, uncertainty aversion, and accountability concerns already in place. Emerging empirical evidence is consistent with this inference. Serious cases of delusional spiraling in users interacting with AI chatbots have been linked to at least fourteen deaths and multiple wrongful-death lawsuits against AI companies [83]. Journalistic accounts describe users who began with curiosity or only tentative beliefs and, through extended interaction, developed unshakable convictions that their chatbot was conscious, that they had made revolutionary scientific discoveries, or that they were receiving personalized supernatural guidance [84–86]. In several reported cases, users eventually suspected that the chatbot was being sycophantic and yet continued to engage with it, unable to exit the spiral [81,87]. Awareness, again, proved insufficient [25].

What makes this loop especially dangerous is that it exploits both motivational and cognitive vulnerabilities. Users do not merely seek information; often, they seek reassurance, legitimation, and relief from uncertainty. In that sense, sycophancy supplies exactly what motivated reasoning demands: support that feels external, authoritative, and objective. A user who asks an LLM whether a business plan is viable, a political position is justified, or a symptom is probably benign may feel that the answer comes from an independent intelligence. In reality, the desired conclusion is being pursued through a channel structurally disposed to accommodate it. The validating response is then experienced not as a reflection of the user's own framing, but as external confirmation. This is artificial confidence in one of its clearest forms. Cheng et al. [14] further showed that sycophantic AI interactions can reduce prosocial intentions and increase dependency, suggesting that the loop may affect not only belief formation but broader patterns of self-regulation and social behavior.

This loop does not operate only within a single exchange. It also exists at the level of the broader system. Contemporary LLMs are trained on vast corpora of human-generated text that increasingly include AI-assisted and AI-generated material. As such content proliferates across the internet, scientific communication, and organizational life, future models are trained on outputs already shaped by the agreeableness, fluency, and validation pressures instilled by RLHF [88]. A recursive dynamic follows: models generate biased or overly validating text; that text re-enters the training environment; later models inherit and may amplify the same tendencies. What is sometimes called model collapse [88] is therefore not only a technical problem. It is a systemic one. Over time, the mirror reflects not just the current user but also the aggregated biases, preferences, and avoidance strategies of prior users.

Systemically, the feedback loop operates across multiple levels and timescales. At the micro level, within a single conversation, confirmatory prompts and validating outputs can generate escalating certainty. At the meso level, repeated interaction patterns may entrench user reliance and reduce engagement with independent sources of scrutiny. Dohnány et al. [89] describe this dynamic as technological folie à deux, invoking the psychiatric phenomenon of a shared delusion maintained by interacting agents. At the macro level, aggregated user behavior feeds back into model training and product design, creating systems increasingly optimized to validate, reassure, and retain engagement. Each level reinforces the others. The cumulative result is not the broadening of judgment promised by optimistic accounts of AI, but the narrowing of epistemic horizons under conditions that feel low-cost, efficient, and objective. This systemic dynamic leads directly to the next problem: the information paradox, or why more available information may worsen rather than improve judgment.

5. The Information Paradox: Why More Data Makes Us Worse

One of the most deeply held assumptions in both public discourse and classical decision theory is that more information leads to better decisions. The assumption has an impressive pedigree. Bayesian accounts of belief revision treat each new piece of evidence as an opportunity to update priors toward the truth [90,91]. On this view, a decision-maker who receives more diagnostic information should converge on more accurate beliefs. Simon's [18] foundational analysis of bounded rationality reinforced the same intuition from the opposite direction: if suboptimal judgment reflects limits in memory, attention, and computational capacity, then relaxing those limits should improve performance. Subsequent work in the heuristics-and-biases tradition [7,9] and in process-tracing research on decision strategies [92] similarly treated cognitive economy as a response to scarcity. When search is costly and memory is limited, people rely on shortcuts that trade accuracy for tractability.

The shared implication across these traditions is clear: if the friction of search were removed, relevant information could be retrieved instantly, organized coherently, and presented fluently, judgment quality should improve. This logic underlies much of the optimism surrounding LLMs. By democratizing access to vast stores of knowledge and eliminating many of the costs once associated with information search, these systems should, in principle, help people think better, decide more wisely, and err less often.

This expectation is not entirely misguided. Under certain conditions, it is correct. When the task is well-defined, the information is genuinely diagnostic, and the decision-maker is motivated primarily by accuracy, additional information can improve performance [22,55]. A physician who gains access to a critical lab result that was previously unavailable may revise a diagnosis in a better direction. A financial analyst who receives a timely earnings report may adjust a forecast more precisely. In such cases, the value of information is straightforward: it fills a real gap in knowledge, and that knowledge is incorporated in a roughly Bayesian fashion.

But these conditions are narrower than they first appear. The classical prediction depends on a further, usually implicit, premise: that decision-makers will attend to relevant information impartially, weigh it in proportion to its diagnostic value, and integrate it into a balanced assessment of the evidence [93]. This is precisely the premise that decades of research in behavioral decision science, social psychology, and political cognition have rendered empirically fragile [9,94,95]. The result is a paradox with two faces. For most users, more information fails to improve judgment because it is filtered through prior beliefs, identity commitments, and motivated interpretation [1,3]. For a smaller but still consequential group, information is not merely misused but actively avoided because the source itself is regarded with suspicion [96]. Both responses produce worse decisions than classical models would predict, and both are driven not by simple deficits in cognitive capacity but by behavioral and emotional tendencies that the LLM interface often fails to correct and may actively amplify.

This matters because LLMs are not neutral repositories of information. They are socio-technical systems that reduce search costs, reorganize epistemic authority, and alter how users encounter, interpret, and deploy evidence. Their promise rests on the assumption that easier access to more information will improve judgment. Their danger lies in the fact that, under ordinary conditions of human use, the same abundance may instead make motivated reasoning easier, faster, and more convincing.

5.1. More Information, Worse Judgment: The Selective-Processing Failure

Lord et al. [20] presented participants who held strong views on capital punishment with two fictitious studies, one supporting and one opposing the deterrent effect of the death penalty. The materials were carefully counterbalanced so that each methodology was equally associated with both pro- and anti-capital-punishment conclusions. Balanced evidence should have produced moderation. It did not. Both proponents and opponents rated the study that confirmed their prior belief as

methodologically superior, dismissed the disconfirming study as flawed, and reported stronger confidence in their original position. The same body of evidence pushed the two groups further apart. Lord et al. termed this process biased assimilation: the tendency to accept confirming evidence with relatively little scrutiny while applying disproportionate skepticism to disconfirming evidence. The pattern has since been replicated across domains such as affirmative action, gun control, environmental policy, and public health [21,75,97].

Once evidence is evaluated selectively, providing more of it no longer guarantees better judgment. It may simply supply additional raw material for selective processing. Taber and Lodge [21] demonstrated this directly. In studies of attitudes toward affirmative action and gun control, participants showed both confirmation and disconfirmation biases: they preferred attitude-consistent sources, counterargued against incongruent evidence, and bolstered congruent evidence with relatively little resistance. Importantly, these effects were strongest among the most politically knowledgeable participants. Those with the greatest informational resources were also the most effective at deploying them in defense of prior commitments. Informational sophistication did not correct bias. It sharpened it.

Kahan et al. [75] extended this logic to science comprehension. In a study of perceived climate-change risks, the science-comprehension thesis predicted that greater scientific literacy and technical reasoning would be associated with greater concern about climate change, as better-informed individuals aligned their views more closely with the scientific consensus. The cultural-cognition thesis predicted instead that individuals would interpret the issue through their cultural worldviews, and that this tendency would be strongest among the most scientifically literate. The results strongly supported the latter account. Among individuals with the highest levels of scientific literacy and numeracy, cultural polarization was greatest: hierarchical individualists with high scientific literacy were less concerned about climate change than those with low scientific literacy, whereas egalitarian communarians with high scientific literacy were more concerned. Science literacy did not mitigate polarization. It intensified it.

If human judgment does not reliably improve with more information, and if the best-informed individuals are sometimes the most biased, then the informational abundance provided by LLMs cannot be assumed to be corrective. The central problem is not simply a lack of access to good information. Under many conditions, it is a lack of motivation or discipline to use that information in the service of accuracy rather than prior commitment. When users approach an LLM with a preferred conclusion already in mind, the model's informational richness does not necessarily serve as a corrective force. It becomes a reservoir from which confirming evidence can be efficiently extracted. The system's ability to retrieve relevant, well-articulated, and seemingly authoritative content does not overcome biased assimilation. Under ordinary conditions, it can accelerate it by delivering precisely the sort of confirming material that users would otherwise have had to search much harder to find.

This logic intersects with the broader literature on information and communication technologies and political polarization. Bail et al. [30] found that exposure to opposing political views on Twitter increased rather than decreased polarization, particularly among Republicans. Madsen et al. [32] likewise showed that congenial political content strengthened partisan identity, whereas cross-cutting content had weaker and less durable effects. Across multiple studies, information environments that expand access to ideologically diverse material do not reliably produce moderation, as users selectively attend to, engage with, and interpret content through the lens of their prior commitments [31,98]. The core mechanism is motivational filtering: the same abundance that could, in principle, provide corrective evidence also creates more opportunity for selective exposure, selective elaboration, and selective dismissal.

LLMs add a distinctive and arguably more potent layer to this dynamic. Unlike social media platforms, which present streams of content for users to sort through, LLMs respond directly to user queries and are optimized for user satisfaction. The result is not merely a biased information environment but an actively responsive one: a system that generates bespoke confirming content in

real time, calibrated to the user's prompt and delivered with the fluency that triggers automatic credibility attributions (see Section 3). Where social-media algorithms aggregate behavioral signals to surface congenial content, LLMs synthesize it on demand, in the precise frame, tone, and register implied by the query [12,79]. In that sense, the information paradox does not disappear in the age of AI. It becomes more interactive, more frictionless, and potentially more severe.

5.2. *Less Information, Worse Judgment: The Aversion Failure*

Importantly, this account captures only part of the picture. To portray the problem as universal over-reliance on LLMs would misread the empirical landscape. Many potential users do not engage with AI systems uncritically; they avoid them altogether. A 2024 U.S. survey reported that 37% of respondents had never used AI tools, while cautious, concerned, and skeptical attitudes toward AI were widespread across the population [99]. Mahmud et al.'s [100] systematic review of the algorithm-aversion literature documents a robust and replicated pattern: across domains ranging from medical diagnosis to financial forecasting to consumer recommendation, people often prefer human judgment to algorithmic judgment even when the algorithm is demonstrably more accurate. This aversion is especially pronounced in domains perceived as consequential or as requiring uniquely human capacities such as empathy or moral reasoning [101,102]. Importantly, Dietvorst et al. [103] showed that observing an algorithm err, even when its overall performance still exceeds that of a human alternative, is enough to produce enduring reluctance to rely on it, whereas comparable human errors do not generate the same reaction.

The behavioral mechanisms underlying aversion are, in many respects, the mirror image of those underlying over-reliance, and they are no less rooted in emotion and motivation. Frenkenberg and Hochman [10] identified two psychologically distinct dimensions of AI anxiety: anticipatory anxiety, a future-oriented apprehension about AI's disruptive consequences for employment, autonomy, and social order, and annihilation anxiety, a deeper existential concern about the erosion of human distinctiveness in the face of increasingly capable machines. Both dimensions correlated significantly with overall AI anxiety, and both predicted avoidance tendencies. Critically, AI anxiety was strongly negatively correlated with use motives ($r = -0.40$), indicating that affective resistance directly suppresses the perceived value of engaging with the technology. The relationship between anxiety and use was also non-linear: anxiety was highest at the extremes of low and high usage and lowest at moderate engagement, suggesting that both unfamiliarity and over-reliance generate emotional discomfort, whereas balanced use may mitigate it. These findings echo earlier work on technostress and computer anxiety [104,105], while extending it to systems whose conversational fluency and apparent autonomy evoke stronger reactions than earlier digital tools did.

The decision-making costs of aversion are not merely the reverse side of over-reliance; they are independently consequential. A clinician who refuses to consult an AI-assisted diagnostic system that might have flagged a rare presentation forgoes potential diagnostic benefit just as surely as one who defers uncritically to a flawed recommendation [29]. A policy analyst who dismisses model output because it conflicts with intuition loses access to potentially disconfirming evidence that could have corrected an initial misjudgment. An investor who abandons algorithmic forecasting after a single visible error forfeits the long-run advantage of statistical prediction over clinical judgment, a pattern documented since Meehl [106] and repeatedly confirmed [107]. In each case, the user's emotional response to the technology—*anxiety, distrust, or perceived threat to professional identity*—substitutes for an evaluation of its diagnostic value, and judgment suffers accordingly.

Aversion and over-reliance, then, are not opposite virtues and vices, but two expressions of the same underlying problem within the human-AI system. In both cases, the user's relation to the tool is governed less by a calibrated assessment of when and how it is informative than by motivational and affective dispositions. The over-reliant user treats the model's output as an authority to defer to; the avoiding user treats it as a threat to reject. Neither engages the information on its merits. In that sense, the central failure is not simply intellectual but behavioral, a point that accounts for human-AI interaction focused only on cognitive capacity are likely to miss.

5.3. The Informational-Relevance Framework

The informational-relevance framework [22,53,108] provides a theoretical lens through which both failure modes can be understood within a single account. On this view, decision quality depends not on the sheer volume of information available, nor even on its objective accuracy in the abstract, but on its relevance: the extent to which the information attended to is diagnostic of the judgment at hand, given the decision maker's orientation and the structural demands of the task. Extending the ecological-rationality tradition [23,109], the framework emphasizes that the match between the informational environment and the user's mode of engagement is a stronger determinant of judgment quality than either factor alone. Relevant, diagnostic information can improve performance when the user is oriented toward accuracy; irrelevant, selectively filtered, or motivationally favored information degrades it, regardless of how much of it is available.

Within this framework, more information improves judgment only when it increases the proportion of task-relevant content that the decision maker actually attends to and integrates. When additional information is neutral, its effect may be negligible. When it is selectively aligned with prior beliefs, its effect can be distortive because it inflates the apparent evidential basis for the prior position without genuinely expanding the user's epistemic horizon [108]. And when the user disengages from the information altogether due to anxiety, distrust, or identity threat, the system's accuracy becomes functionally irrelevant because none of it enters the decision-making process.

Applied to the human-AI interface, this framework highlights the central role of informational behavior within a socio-technical system. A user who asks an LLM balanced, open-ended questions and weighs the response against alternative sources may receive output that is informationally relevant and diagnostically useful. A user who asks a leading question filters the system's vast informational resources through a narrow aperture, producing output that is informationally abundant yet diagnostically impoverished. A user who refuses to consult the system at all forfeits whatever relevant information it might have provided. The system itself does not reliably distinguish among these orientations; it responds to leading prompts with the same fluency it brings to balanced ones, and it offers nothing to the user who never queries it. The critical difference, therefore, lies not only in the model's architecture but in the user's epistemic and affective orientation toward it.

This perspective also helps explain why the standard remedies proposed for the information paradox—more education, greater transparency, and broader data access—are unlikely to resolve it on their own. Education may increase users' ability to process information, but informational skill can just as easily increase the efficiency of motivated reasoning as constrain it [19,21]. Transparency in AI outputs, such as confidence indicators, citations, or caveats, may improve calibration, but only if users are motivated to attend to those signals rather than to the confirming substance of the response itself. Greater data access is valuable only insofar as the data accessed is diagnostically relevant, which depends as much on user goals as on system capability. And none of these interventions directly addresses the avoiding user, for whom the primary barrier is not informational but affective. Interventions aimed at improving human-AI judgment must therefore target not only the information provided, but also the user's informational conduct: how prompts are formulated, how outputs are evaluated, how emotional reactions to the technology are managed, and what motivational orientation governs the interaction as a whole.

From this perspective, the information paradox is not a failure of technology alone, but a predictable consequence of deploying powerful informational tools within a psychological architecture oriented, at different times, toward accuracy, coherence, comfort, confirmation, or threat avoidance. When a telescope extends vision, the user has little influence over what the instrument reveals. When an LLM extends cognition, the user helps shape the observation from the outset through prompt framing, topic selection, affective stance, and even the prior decision to query the system at all. The telescope is constrained by optics; the LLM is constrained by statistics, optimization, and the user's preferences and aversions. The cognitive extension provided by LLMs is therefore qualitatively different from that provided by earlier tools: it amplifies not only capacity, but disposition. For users oriented toward accuracy, that amplification can be genuinely beneficial. For

users oriented toward confirmation, it becomes a refined form of self-validation. For users oriented toward avoidance, it yields nothing at all. The outcome is not determined by information quality alone in any of these cases. The paradox is that the tool that promises unprecedented access to knowledge may, depending on who uses it and how, also produce a sophisticated form of ignorance masked by fluency, certainty, and the appearance of epistemic authority. This transformation of information abundance produces not enlightenment but an artificial confidence, which the next section addresses.

6. Artificial Confidence

This section introduces *artificial confidence* as the integrative concept that captures the cumulative outcome of the converging processes documented throughout this paper. Artificial confidence refers to a state in which a user's subjective certainty about the content of an AI-mediated exchange is sustained by the structural features of the interaction itself: the fluency of the response, the perceived authority of the source, the absence of friction or contradiction, and the social and commercial logic of a system designed to keep the user engaged. It is *artificial* in two senses. First, it is generated by an artifact rather than earned through inquiry. Second, it is confidence that does not track the evidential warrant for the belief it concerns. Crucially, artificial confidence is not a mild case of poor calibration, in which people believe they know more than they do [36]. It is a categorically different phenomenon in which an entire ecosystem—users, developers, institutions, and arguably the systems themselves—converges on the implicit premise that, for practical purposes, the machine is almost always right.

6.1. Beyond Calibration: The Difference Between Artificial Confidence and Overconfidence

The classical literature on overconfidence locates the problem within the individual decision maker. Overconfidence is typically defined as a mismatch between expressed certainty and actual accuracy, arising from faulty probability estimation, insensitivity to feedback, or motivated self-enhancement [1,36,110,111]. Thus, remedies focused on better feedback, training, and reflective deliberation, based on the assumption that users generally understand the sources of their beliefs but overestimate the precision of their own judgments.

Artificial confidence is not simply overconfidence intensified. The problem is not that users misjudge their own competence, but that they come to treat system outputs as sufficiently trustworthy to act upon without verification. In ordinary human exchange, recipients routinely evaluate testimony by asking who is speaking, how reliable the source is, and what motives may be shaping the message [112]. Much of that epistemic scrutiny may be suspended when the source is a fluent and apparently authoritative LLM [68,113]. The result is not a modest calibration error layered onto an otherwise intact epistemic relationship, but a structural shift in that relationship: the normal cognitive defenses against unreliable testimony are weakened before they fully engage.

Three features distinguish artificial confidence from classical overconfidence. First, it is relational rather than intrapersonal: it emerges at the interface between the user and the system, and neither side can fully explain or correct it on its own. Second, it is socially and systemically reinforced: users, developers, and institutions all help normalize the expectation that these outputs are usually safe to trust, even when formal disclaimers suggest caution. Third, it is self-concealing: the same fluency, coherence, and apparent authority that inflate certainty also suppress the cues that would ordinarily trigger skepticism. Classical overconfidence can, at least in principle, be corrected by clear feedback. Artificial confidence is harder to recalibrate because the user often never encounters the immediate disconfirmation that would expose the absence of an external warrant. A person who consults a chatbot, receives a confident answer, and acts on it may never confront the crucial epistemic fact that the belief felt validated without ever being independently verified [114].

6.2. The Erosion of Epistemic Vigilance

Humans are not passive absorbers of information. According to Sperber et al. [112], people possess a set of cognitive mechanisms—source monitoring, plausibility checking, coherence evaluation, and social inference about a speaker’s competence and motives—that filter testimony before it is accepted as belief. Epistemic vigilance protects against deception, manipulation, and sincere but mistaken testimony. It is not an elite skill possessed by only some users, but a baseline feature of ordinary communication. Crucially, it calibrates trust rather than eliminating it: it helps people decide when, how much, and on what basis to believe.

Conversational AI may erode that vigilance through several converging features of the interface. First, the cues that ordinarily activate it—clear source attribution, visible expertise, reputational history, and signs that a speaker has interests at stake—are largely absent. Within the Computers Are Social Actors framework [115,116], users readily apply social heuristics to machine interlocutors, yet the system provides few of the evaluative cues that usually accompany human testimony. It has no biography, no visible track record, no audible hesitation, and no discernible stake in the claim. It appears, phenomenologically, as an oracle without provenance. Second, its outputs are optimized for fluency and confidence, both of which increase perceived credibility irrespective of truth [69,71,72]. Third, the interface presents heterogeneous training material as though it were the voice of a single, comprehensive, reliable informant, masking the variation in quality, perspective, and evidential status of the materials from which the response is statistically assembled.

The result is that epistemic vigilance is not eliminated but relaxed precisely where it is most needed. Metzger and Flanagin [117] showed that users evaluating online information rely heavily on cognitive heuristics rather than substantive scrutiny, especially when interfaces project authority. With LLMs, those credibility cues are not merely added to the content; they are built into the mode of delivery itself. Claims arrive in the same polished register and with the same confidence regardless of their epistemic status. Users are therefore left with fewer metacognitive footholds for skeptical engagement. Recent empirical work directly supports this concern. Steyvers et al. [113] identified a systematic calibration gap between users’ confidence in LLM answers and the models’ own internal confidence: users overestimated accuracy, relying on surface cues such as response length and explanatory fluency rather than evidential substance. Fernandes et al. [118] extended this pattern, showing that people can perform better with AI assistance while simultaneously becoming less accurate judges of the quality of that performance. More strikingly, greater AI literacy did not yield more discriminating use, but more inflated self-assessment. Familiarity with the tool, in other words, may strengthen confidence faster than it strengthens scrutiny. The erosion of epistemic vigilance is therefore not merely a user weakness. It is a predictable feature of a sociotechnical arrangement that concentrates authority, removes provenance, and packages heterogeneous information as frictionless conversational certainty.

6.3. *The Convergence of User, Developer, and System on the Premise of Near-Perfection*

Artificial confidence is not a private cognitive error attributable to naïve users. It is sustained by a convergence of expectations across the entire sociotechnical system. Users approach these systems with the implicit assumption that they are extraordinarily competent. Developers and the firms that deploy them describe them in terms that, even when carefully hedged, position the technology as a transformative cognitive resource. Institutional adopters then embed these systems into workflows on the assumption that their outputs are reliable by default. And the systems themselves present a stance of competent informativeness that users have little reason to question.

This convergence is especially visible in the disclaimers that accompany contemporary AI systems. Most major commercial LLMs display warnings stating that the system may make mistakes and that important information should be verified. Yet these warnings are typically rendered in small, persistent, low-salience formats and are plausibly subject to the same habituation dynamics documented for security warnings, browser alerts, and terms-of-service notices, where repeated exposure to low-consequence prompts reduces attention and compliance over time [119–121]. The problem, however, is not merely that users ignore the warnings. It is that the warnings may

themselves communicate a meta-message that undercuts their literal content. By stating that the system “may” make mistakes—rather than that it routinely does, that many of its errors are difficult for ordinary users to detect, and that its fluency is optimized through training procedures not directly tied to factual accuracy [43,122]—the disclaimer is easily read as implying that error is exceptional against a baseline of reliability. The pragmatic implicature is that verification is a precautionary measure reserved for unusually high-stakes cases, not a baseline epistemic requirement in ordinary use. In this way, the disclaimer acknowledges fallibility while simultaneously normalizing the assumption that fallibility is rare.

Even well-intentioned expressions of professional humility can have a similar structure when they become formulaic rather than calibrated to the evidential limits of a particular case [37,39]. Clinicians who present every diagnosis with uniform confidence are not communicating meaningful uncertainty; they are performing a ritual acknowledgment whose routinization signals that uncertainty is merely a generic background condition rather than a live feature of the present judgment. The disclaimer attached to LLM interfaces operates much the same way. It allows both user and developer to preserve the comfortable fiction that responsibility for verification has been formally transferred to the user, even as the actual structure of the interaction makes such verification psychologically and practically unlikely.

The result is what may be called a distributed epistemic asymmetry: every actor can plausibly claim to have acknowledged the system’s limitations, yet no actor behaves as though those limitations are operative. Developers cite the disclaimer; users cite the developer’s expertise; institutions cite both; and the system, insofar as the term “behavior” can be applied to a stochastic generator, produces fluent and confident outputs whether the situation warrants them or not. At the level of the ecosystem, artificial confidence is the equilibrium state of this convergence: a condition in which everyone has, in principle, signed off on the proposition that the system is fallible, yet proceeds on the operative assumption that it is not.

6.4. *From Episodic Error to Dispositional Erosion*

Sections 4 and 5 traced the local mechanisms by which sycophantic responses validate user priors and informational abundance fails to produce calibration. The stronger claim advanced here is that artificial confidence may do more than generate isolated epistemic errors. It may gradually shift the user’s epistemic posture toward a generalized disengagement from the practices that constitute careful inquiry. That shift forms the bridge between the micro-level dynamics of single interactions and the institutional pathologies considered in the next section.

Three observations support this stronger reading. First, the disengagement appears to outlast the encounter that produced it. Frenkenberg and Hochman [10] show that as users become more familiar with AI tools, dependency increases while critical engagement does not increase in parallel. Research on trust development in human–AI teams points in the same direction: trust grows with repeated interaction but does not reliably track performance [123]. What is being reinforced, in other words, is not a calibrated ability to distinguish when the tool can be trusted from when it cannot, but a generalized disposition toward reliance.

Second, the disengagement reshapes expectations about what an informational exchange should feel like. Cheng et al. [14] report that interactions with sycophantic AI systems decrease prosocial intentions and increase dependence. This is consistent with the broader claim that frictionless agreement recalibrates the affective baseline of inquiry itself. Once users become accustomed to fluent, validating, instantly available answers, the more demanding work of independent verification, comparison across sources, and tolerance of unresolved uncertainty begins to feel disproportionately effortful and unrewarding. The relevant contrast is no longer between informed and uninformed judgment, but between an easy, seemingly “accurate” answer and a difficult, potentially fallible inquiry. Under that contrast, and given the well-documented human preference for cognitive economy [7,124], the former is likely to win by default.

Third, and most consequentially, this is precisely the kind of shift that the concept of epistemic humility, in its dispositional sense, is meant to forestall [37,125]. Epistemic humility, as developed in contemporary virtue epistemology, is not an episodic acknowledgment that one might be wrong about a particular claim. It is a standing readiness to treat one's own beliefs as provisional and to remain open to the kinds of friction—disagreement, recalcitrant evidence, alternative framings—through which beliefs become revisable [126–128]. On this view, the disposition is partly constituted by the practices that expose beliefs to friction. Remove the friction, and what is eroded is not merely the occasional act of revision, but the standing readiness that makes revision possible. Artificial confidence, therefore, threatens epistemic humility not primarily by producing specific false beliefs that resist correction, but by removing the very conditions that sustain revisability. A user whose informational environment systematically delivers smooth, confirming, authoritative responses is not simply a user whose calibrated beliefs happen to be wrong. It is a user who has gradually lost the practice of holding beliefs in a way that permits revision.

The point, then, is not that LLMs add one more bias to an already familiar list. It is that, by virtue of their architecture and the affective texture of the interactions they enable, they may erode the dispositional substrate on which corrective practice depends.

6.5. *From Individual Disposition to Institutional Practice*

If artificial confidence operates not only at the level of particular beliefs but also at the dispositional level of inquiry itself, its consequences cannot remain confined to individual users. They diffuse through the professions, politics, and organizations in which those users are embedded. The deeper concern, then, is not the occurrence of any single error, but a structural shift in the conditions under which errors are produced, recognized, and corrected.

In clinical medicine, this dynamic compounds the well-documented risks of automation bias and diagnostic overconfidence [28,29,38,66]. A clinician frames the question, the system narrows the field of plausible answers, the fluency of the response suppresses alternatives, and the resulting diagnostic certainty may exceed its evidential basis in ways that neither clinician nor patient can easily detect. The danger is not only deference to incorrect recommendations. It is the gradual erosion of diagnostic habits—keeping the differential open, seeking disconfirming evidence, and communicating uncertainty honestly—that distinguishes careful clinical reasoning from procedural pattern-matching [37,129].

In political and public discourse, artificial confidence may intensify the polarization dynamics already documented in pre-LLM environments [19,21,30]. LLMs can generate bespoke confirming arguments in real time, calibrated to the user's framing and articulated in the register of professional analysis. The user's resulting confidence is no longer bounded by the limits of personal argumentative imagination, but extended by the system's generative capacity. The collective consequence is not merely individual overconfidence. It is epistemic fragmentation: opposing groups acquire AI-validated certainty in mutually incompatible claims, without meaningful exposure to the strongest form of the opposing position.

In organizational settings, artificial confidence facilitates a distinctive diffusion of responsibility. Grote and Berens [33] argue that machine-learning systems in healthcare generate trade-offs at both epistemic and normative levels, potentially undermining clinicians' epistemic authority while diffusing moral responsibility across opaque sociotechnical systems. The present analysis suggests that this dynamic extends well beyond medicine. In any institutional setting where AI tools inform decisions—hiring, strategy, risk assessment, compliance—the availability of an apparently authoritative output shifts accountability away from the decision-maker and toward the tool, even when the output itself was shaped by the decision-maker's framing of the query [4,5,130]. The individual who can point to an AI recommendation feels less personally accountable for the outcome; the organization that can describe its decisions as “data-driven” acquires a rhetorical shield that may obscure the motivated, selective, and human-directed process by which the data were queried and the recommendation produced.

A particularly instructive version of this institutional dynamic appears in universities and academic publishing. Two responses have predominated, and both fail in characteristic ways. The first is categorical prohibition: blanket bans on LLM use in coursework, manuscript preparation, or peer review. Such bans are not only difficult to enforce; they also collapse a heterogeneous set of practices with very different epistemic implications into a single prohibited object. Asking an LLM to surface possible objections to a draft argument is not the same activity as outsourcing the framing of one's contribution to it, and a policy that does not distinguish between the two cannot guide responsible use. The second response is the opposite extreme: tacit normalization, in which institutions decline to articulate meaningful guidance and leave individual scholars to negotiate the boundary between assistance and ghost-authorship on their own. Both responses avoid the same underlying question: how the tool reshapes the cognitive practices of its users.

A more promising alternative has begun to emerge in recent editorial guidance. Baer and Kouchaki [131], in a governance framework for *Organizational Behavior and Human Decision Processes*, locate the problem in the right place: not in model accuracy alone, but in the temptation to let fluent, seemingly insightful output substitute for primary-source engagement and independent reasoning. Their framework distinguishes supportive uses, such as literature search, code debugging, and copy-editing, from substantive uses that affect a work's intellectual contribution, and it requires that authors retain full responsibility for verification throughout. Its significance lies not in solving the underlying psychological problem, but in identifying its locus correctly. The core issue is not whether the technology is becoming more accurate. It is whether users are allowed to substitute apparent insight for inquiry.

How the problem is framed determines the form of the response. A field that treats LLM use as a purely technical issue of hallucination rates, detection tools, or output accuracy will tend to produce policies aimed at sorting AI-assisted work from human work. A field that treats it as a systems problem in the preservation of inquiry within human-AI environments can instead develop policies aimed at protecting the practices that make scholarship and professional judgment possible: primary-source engagement, independent verification, accountable reasoning, and openness to revision. Researchers in judgment and decision-making are especially well-positioned, and arguably especially obligated, to push the institutional response in that direction. The evidence reviewed throughout this paper on automation bias, sycophancy, motivated reasoning, the metacognitive disconnect [113,118], and the dependency dynamics [10] provide a basis for evidence-informed policy that neither blanket prohibition nor tacit normalization can match.

Across domains, the deeper concern is the same: professions and organizations that lose the practice of epistemic vigilance do not merely make more mistakes. They may also lose the cognitive infrastructure through which mistakes are recognized, contested, and corrected. This, ultimately, is the institutional danger captured by the concept of artificial confidence, and it is the danger from which the concluding section proceeds.

7. Conclusions

This paper has advanced a central claim: the primary risk posed by LLMs is not that they sometimes produce incorrect answers, but that they can make users feel more certain without becoming more warranted in their beliefs. Their danger lies less in isolated technical failure than in the cognitive and affective conditions under which their outputs are received. Users do not approach LLMs as neutral truth-seekers. They approach them with prior beliefs, directional goals, varying tolerance for uncertainty, and a preference for fluent, low-friction cognition. LLMs do not bypass this architecture. They engage it with unprecedented efficiency.

The argument developed here provides an explanation. Because LLM outputs are highly sensitive to user framing, the apparent objectivity of the response often obscures its dependence on the prompt's subjectivity. Because these systems are optimized for fluency, responsiveness, and user satisfaction, they can amplify familiar distortions in human judgment rather than correct them. And because greater informational abundance does not reliably produce more balanced information

processing, access to more data can intensify motivated reasoning, selective confirmation, and unwarranted certainty rather than reduce them. The cumulative result is what this paper has called *artificial confidence*: a relational, structurally reinforced, and often self-concealing form of certainty sustained by the interaction itself rather than by the quality of the evidence.

This analysis also carries an important implication about the locus of intervention. If the central problem were primarily technical (e.g., hallucinations, factual error, or training-data imperfections), then improvements in model accuracy might be sufficient. But the argument here suggests that technical refinement alone will not resolve the underlying concern. A more accurate model may still be used for confirmation-seeking. A more fluent model may still encourage epistemic passivity. A more authoritative model may still provide rhetorical cover for users and institutions seeking to offload responsibility. The problem is not exhausted by what the system outputs. It lies in the relationship users, developers, and institutions build around those outputs.

That is why the stakes are broader than isolated mistakes. At the individual level, artificial confidence may erode epistemic vigilance while appearing to expand competence. At the professional level, it may displace the practices that define responsible judgment: seeking disconfirming evidence, holding hypotheses open, engaging primary sources, and tolerating uncertainty when uncertainty is warranted [37,131]. At the collective level, it may deepen epistemic fragmentation by allowing opposing groups to acquire AI-mediated certainty in incompatible claims. In that sense, the cumulative concern is not that LLMs will make humans less intelligent, but that they may make humans less corrigible.

The practical implication is therefore not that such tools should be abandoned, but that their integration must be governed by norms and designs that preserve revisability. For users, that means treating outputs as hypotheses rather than verdicts. For developers, it means recognizing that frictionless agreement and low-salience disclaimers may themselves be sources of epistemic harm. For institutions, it means protecting the practices of inquiry that must remain intact regardless of which tools are used. For researchers in judgment and decision-making, it means taking a more active role in shaping this discussion, because the core questions raised by LLM use are, ultimately, questions about human judgment.

If artificial confidence is the risk, then epistemic humility is not a moral luxury but a practical necessity. Whether AI becomes an extension of human intelligence or a substitute for disciplined inquiry will depend less on the sophistication of the machine than on whether users and institutions preserve the conditions under which beliefs remain open to revision.

Beyond psychology, artificial confidence is also a governance problem. If AI-augmented systems are to be used responsibly, organizations cannot assume that better models alone will protect decision quality. They must also design oversight structures, verification routines, and decision protocols that account for predictable human tendencies toward overreliance, confirmation-seeking, and responsibility offloading. In this sense, the challenge is not only to improve artificial intelligence but to build socio-technical systems in which human judgment remains accountable, interruptible, and epistemically engaged.

7.1. Limitations and Future Directions of Research

The present paper is interpretive and integrative rather than empirical, and its limitations should be acknowledged. The concept of artificial confidence, as developed here, is a theoretical construct synthesized from findings across judgment and decision-making, human-computer interaction, virtue epistemology, clinical reasoning, and political psychology. Although the component literatures are empirical, the broader claim that these processes converge on a distinct and practically consequential state that erodes the dispositional basis of epistemic humility remains a hypothesis that requires direct testing. A clear priority for future research is therefore to operationalize artificial confidence in ways that distinguish it empirically from classical overconfidence, including its correlates, behavioral signatures, and resistance to correction.

A second limitation concerns the recency and instability of the empirical literature on human–LLM interaction. Much of the relevant evidence has emerged only in the last few years, and both the technology and its modes of use are changing rapidly. Findings derived from the current generation of models may not generalize straightforwardly to systems with different architectures, training procedures, or interface designs. Longitudinal work tracking users across successive model generations would therefore be especially valuable, as would experimental designs that manipulate interface features such as fluency, contradiction, uncertainty displays, and prompt structure.

Third, the analysis here has focused mainly on text-based conversational AI. Multimodal systems, agentic systems that act on the user’s behalf, and AI embedded in specialized professional workflows may raise related but not identical risks. The broader argument advanced here may extend to such contexts, but the relevant mechanisms are likely to differ. Future work should therefore examine whether artificial confidence is expressed differently when the system generates images, takes actions, triages information, or operates under domain-specific institutional constraints.

Fourth, the present analysis has largely bracketed individual differences. Although the paper has emphasized common vulnerabilities in human judgment, susceptibility to artificial confidence is unlikely to be uniform. Need for cognition, need for closure, dispositional epistemic humility, domain expertise, trust in AI, and prior ideological commitments are all plausible moderators (Krumrei-Mancuso & Rouse, 2016; Kruglanski & Webster, 1996). Identifying which users are most vulnerable, under which task conditions, and with which protective supports should be a central goal of future work.

Finally, this paper has deliberately emphasized risk. That emphasis reflects the current imbalance in public and scholarly discourse, which remains heavily oriented toward capability and promise. It should not be read as a denial that LLMs can, under the right conditions, support careful inquiry. An equally important complementary agenda is to identify when and how they do so, and what design, educational, and institutional conditions make beneficial use more likely. A balanced literature will need to explain not only how LLMs can erode judgment, but also when they scaffold it—and what separates those two outcomes.

7.2. Summary

The greatest risk posed by LLMs is not that they will replace human judgment, but that they will leave it unchallenged while making it feel more rigorously supported than ever before. A tool that confirms what its user already believes—in the register of expertise, with the fluency of authority, and with minimal friction—does not simply extend cognition. It can also enclose it. The appropriate response is not abandonment, which would be neither realistic nor desirable, but the preservation of the practices that make beliefs revisable: a willingness to be wrong, a discipline of engaging the strongest opposing case, a tolerance for uncertainty when uncertainty is warranted, and a recognition that the comfort of agreement is often a sign that something has gone unexamined. These are among the core practices of epistemic humility, and they are precisely the practices that artificial confidence threatens to erode. Whether AI becomes an extension of human intelligence or a substitute for disciplined inquiry will ultimately be decided by human choice. It will depend on whether users and institutions remain committed to the conditions under which beliefs can still be questioned, revised, and corrected.

Funding: This work received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kunda, Z. The case for motivated reasoning. *Psychol. Bull.* **1990**, 108, 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>

2. Mercier, H.; Sperber, D. Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* **2011**, *34*, 57–74. <https://doi.org/10.1017/S0140525X10000968>
3. Hochman, G. Rationalization as Cognitive Homeostasis: A Homobiasos Theory of Adaptive Self-Regulation. *PsyArxiv* **2025**, preprint. Available online: https://osf.io/w5cqr_v1 (accessed on 13 April 2026).
4. Lerner, J.S.; Tetlock, P.E. Accounting for the effects of accountability. *Psychol. Bull.* **1999**, *125*, 255–275. <https://doi.org/10.1037/0033-2909.125.2.255>
5. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does automation bias decision-making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. <https://doi.org/10.1006/ijhc.1999.0252>
6. Epstein, S. Integration of the cognitive and the psychodynamic unconscious. *Am. Psychol.* **1994**, *49*, 709–724. <https://doi.org/10.1037/0003-066X.49.8.709>
7. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus and Giroux: New York, NY, USA, **2011**.
8. Sloman, S.A. The empirical case for two systems of reasoning. *Psychol. Bull.* **1996**, *119*, 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
9. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
10. Frenkenberg, A.; Hochman, G. It's scary to use it, it's scary to refuse it: The psychological dimensions of AI adoption—anxiety, motives, and dependency. *Systems* **2025**, *13*, 82. <https://doi.org/10.3390/systems13020082>
11. Grassini, S. Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Educ. Sci.* **2023**, *13*, 692. <https://doi.org/10.3390/educsci13070692>
12. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askeell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards understanding sycophancy in language models. *arXiv* **2023**. <https://doi.org/10.48550/arXiv.2310.13548>
13. Ibrahim, L.; Hafner, F.S.; Rocher, L. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2507.21919>
14. Cheng, M.; Lee, C.; Khadpe, P.; Yu, S.; Han, D.; Jurafsky, D. Sycophantic AI decreases prosocial intentions and promotes dependence. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2510.01395>
15. Brunyé, T.T.; Mitroff, S.R.; Elmore, J.G. Artificial intelligence and computer-aided diagnosis in diagnostic decisions: 5 questions for medical informatics and human-computer interface research. *J. Am. Med. Inform. Assoc.* **2025**, advance online publication. <https://doi.org/10.1093/jamia/ocaf123>
16. Echterhoff, J.M.; Liu, Y.; Alessa, A.; McAuley, J.; He, Z. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, **2024**; pp. 12640–12653. <https://aclanthology.org/2024.findings-emnlp.739/>
17. Jussupow, E.; Spohrer, K.; Heinzl, A.; Gawlitza, J. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inf. Syst. Res.* **2021**, *32*, 713–735. <https://doi.org/10.1287/isre.2020.0980>
18. Simon, H.A. Theories of bounded rationality. In *Decision and Organization*; McGuire, C.B., Radner, R., Eds.; North-Holland: Amsterdam, The Netherlands, **1972**; pp. 161–176.
19. Kahan, D.M.; Peters, E.; Wittlin, M.; Slovic, P.; Ouellette, L.L.; Braman, D.; Mandel, G. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nat. Clim. Change* **2012**, *2*, 732–735. <https://doi.org/10.1038/nclimate1547>
20. Lord, C.G.; Ross, L.; Lepper, M.R. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **1979**, *37*, 2098–2109. <https://doi.org/10.1037/0022-3514.37.11.2098>
21. Taber, C.S.; Lodge, M. Motivated skepticism in the evaluation of political beliefs. *Am. J. Polit. Sci.* **2006**, *50*, 755–769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
22. Hochman, G. Beyond the surface: A new perspective on dual-system theories in decision-making. *Behav. Sci.* **2024**, *14*, 1028. <https://doi.org/10.3390/bs14111028>
23. Gigerenzer, G. On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychol. Rev.* **1996**, *103*, 592–596. <https://doi.org/10.1037/0033-295X.103.3.592>

24. Hill, K.; Valentino-DeVries, J. What OpenAI Did When ChatGPT Users Lost Touch With Reality. *The New York Times*, 23 November 2025. Available online: <https://www.nytimes.com/2025/11/23/technology/openai-chatgpt-users-risks.html> (accessed on 13 April 2026).
25. Chandra, K.; Kleiman-Weiner, M.; Ragan-Kelley, J.; Tenenbaum, J.B. Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv* 2026. <https://doi.org/10.48550/arXiv.2602.19141>
26. Hudon, A.; Stip, E. Delusional experiences emerging from AI chatbot interactions or “AI Psychosis”. *JMIR Ment. Health*, 2025, 12(1), e85799. <https://doi.org/10.2196/85799>
27. Bellahsen-Harrar, Y.; Lubrano, M.; Lépine, C.; Beaufrère, A.; Bocciarelli, C.; Brunet, A.; Decroix, E.; El-Sissy, F.N.; Fabiani, B.; Morini, A.; et al. Exploring the risks of over-reliance on AI in diagnostic pathology. What lessons can be learned to support the training of young pathologists? *PLoS ONE* 2025, 20, e0323270. <https://doi.org/10.1371/journal.pone.0323270>
28. Berner, E.S.; Graber, M.L. Overconfidence as a cause of diagnostic error in medicine. *Am. J. Med.* 2008, 121 (Suppl. 5), S2–S23. <https://doi.org/10.1016/j.amjmed.2008.01.001>
29. Tavares, T.; Sangalli, L.; Menon, R.S.; Miller, C.S.; Hochman, G. Educational frameworks for diagnostic decision-making in AI-enhanced head and neck pathology. *Head Neck Pathol.* **in press**.
30. Bail, C.A.; Argyle, L.P.; Brown, T.W.; Bumpus, J.P.; Chen, H.; Hunzaker, M.B.F.; Lee, J.; Mann, M.; Merhout, F.; Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. USA* 2018, 115, 9216–9221. <https://doi.org/10.1073/pnas.1804840115>
31. Cook, J.; Lewandowsky, S. Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Top. Cogn. Sci.* 2016, 8, 160–179. <https://doi.org/10.1111/tops.12186>
32. Madsen, J.K.; Bailey, R.M.; Pilditch, T.D. Large networks of rational agents form persistent echo chambers. *Sci. Rep.* 2018, 8, 12391. <https://doi.org/10.1038/s41598-018-25558-7>
33. Grote, T.; Berens, P. On the ethics of algorithmic decision-making in healthcare. *J. Med. Ethics* 2020, 46, 205–211. <https://doi.org/10.1136/medethics-2019-105586>
34. Patil, S.V.; Myers, C.G.; Lu-Myers, Y. Calibrating AI reliance—A physician’s superhuman dilemma. *JAMA Health Forum* 2025, 6, e250106. <https://doi.org/10.1001/jamahealthforum.2025.0106>
35. Meyer, A.N.D.; Payne, V.L.; Meeks, D.W.; Rao, R.; Singh, H. Physicians’ diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Intern. Med.* 2013, 173, 1952–1958. <https://doi.org/10.1001/jamainternmed.2013.10081>
36. Moore, D.A.; Healy, P.J. The trouble with overconfidence. *Psychol. Rev.* 2008, 115, 502–517. <https://doi.org/10.1037/0033-295X.115.2.502>
37. Hochman, G. Epistemic Humility and Epistemic Confidence: Competing Ethical Forces in Clinical Medicine. 2026, submitted. Available online: https://doi.org/10.31235/osf.io/m6bzt_v1 (accessed on 13 April 2026).
38. Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Med.* 2003, 78, 775–780. <https://doi.org/10.1097/00001888-200308000-00003>
39. Schwab, A. Epistemic humility and medical practice: Translating epistemic categories into ethical obligations. *J. Med. Philos.* 2012, 37, 28–48. <https://doi.org/10.1093/jmp/jhr054>
40. Brynjolfsson, E.; McAfee, A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*; W.W. Norton: New York, NY, USA, 2014.
41. Achiam, J.; Adler, S.; Agarwal, S.; et al. GPT-4 technical report. *arXiv* 2023, arXiv:2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>
42. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 2020, 33, 1877–1901.
43. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 2022, 35, 27730–27744.
44. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30, 5998–6008.

45. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, 3–10 March 2021; pp. 610–623. <https://doi.org/10.1145/3442188.3445922>
46. Shanahan, M. Talking about large language models. *Commun. ACM* **2024**, *67*, 68–79. <https://doi.org/10.1145/3624724>
47. Frankfurt, H.G. *On Bullshit*; Princeton University Press: Princeton, NJ, USA, **2005**.
48. Marcus, G.; Davis, E. *Rebooting AI: Building Artificial Intelligence We Can Trust*; Pantheon Books: New York, NY, USA, **2019**.
49. Feltovich, P.J.; Prietula, M.J.; Ericsson, K.A. Studies of expertise from psychological perspectives. In *The Cambridge Handbook of Expertise and Expert Performance*; Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R., Eds.; Cambridge University Press: Cambridge, UK, **2006**; pp. 41–67.
50. Cheung, V.; Maier, M.; Lieder, F. Large language models show amplified cognitive biases in moral decision-making. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2412015122. <https://doi.org/10.1073/pnas.2412015122>
51. Jung, J.; Lee, H.; Jung, H.; Kim, H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon* **2023**, *9*, e16110. <https://doi.org/10.1016/j.heliyon.2023.e16110>
52. Suri, G.; Slater, L.R.; Ziaee, A.; Nguyen, M. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *J. Exp. Psychol. Gen.* **2024**, *153*, 1066.
53. Ayal, S.; Rusou, Z.R.; Zakay, D.; Hochman, G. Determinants of judgment and decision making quality: The interplay between information processing style and situational factors. *Front. Psychol.* **2015**, *6*, 1088. <https://doi.org/10.3389/fpsyg.2015.01088>
54. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Scialom, T. Augmented language models: A survey. *arXiv* **2023**. <https://doi.org/10.48550/arXiv.2302.07842>
55. Gigerenzer, G.; Gaissmaier, W. Heuristic decision making. *Annu. Rev. Psychol.* **2011**, *62*, 451–482. <https://doi.org/10.1146/annurev-psych-120709-145346>
56. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4299–4307.
57. Malmqvist, L. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing*; Arai, K., Ed.; Springer: Cham, Switzerland, **2025**; pp. 61–74. https://doi.org/10.1007/978-3-031-92611-2_5
58. Wang, C.; Wang, H.; Li, Y.; Dai, J.; Gu, X.; Yu, T. Factors influencing university students' behavioral intention to use generative artificial intelligence: Integrating the theory of planned behavior and AI literacy. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 6649–6671. <https://doi.org/10.1080/10447318.2024.2383033>
59. Clark, A.; Chalmers, D. The extended mind. *Analysis* **1998**, *58*, 7–19. <https://doi.org/10.1093/analys/58.1.7>
60. Heersmink, R. Dimensions of integration in embedded and extended cognitive systems. *Phenomenol. Cogn. Sci.* **2015**, *14*, 577–598. <https://doi.org/10.1007/s11097-014-9355-1>
61. Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **1998**, *2*, 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
62. Findley, K.A.; Scott, M.S. The multiple dimensions of tunnel vision in criminal cases. *Wis. Law Rev.* **2006**, *2006*, 291–397. <https://api.semanticscholar.org/CorpusID:147146926>
63. Mynatt, C.R.; Doherty, M.E.; Tweney, R.D. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Q. J. Exp. Psychol.* **1977**, *29*, 85–95. <https://doi.org/10.1080/00335557743000053>
64. Stanovich, K.E.; West, R.F. Natural myside bias is independent of cognitive ability. *Think. Reason.* **2007**, *13*, 225–247. <https://doi.org/10.1080/13546780600780796>
65. Parasuraman, R.; Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **1997**, *39*, 230–253. <https://doi.org/10.1518/001872097778543886>
66. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
67. Mosier, K.L.; Skitka, L.J.; Heers, S.; Burdick, M. Automation bias: Decision making and performance in high-tech cockpits. *Int. J. Aviat. Psychol.* **1998**, *8*, 47–63. https://doi.org/10.1207/s15327108ijap0801_3

68. Jakesch, M.; Hancock, J.T.; Naaman, M. Human heuristics for AI-generated language are flawed. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2208839120. <https://doi.org/10.1073/pnas.2208839120>
69. Alter, A.L.; Oppenheimer, D.M. Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* **2009**, *13*, 219–235. <https://doi.org/10.1177/1088868309341564>
70. Oppenheimer, D.M. The secret life of fluency. *Trends Cogn. Sci.* **2008**, *12*, 237–241. <https://doi.org/10.1016/j.tics.2008.02.014>
71. Reber, R.; Schwarz, N. Effects of perceptual fluency on judgments of truth. *Conscious. Cogn.* **1999**, *8*, 338–342. <https://doi.org/10.1006/ccog.1999.0386>
72. Fazio, L.K.; Brashier, N.M.; Payne, B.K.; Marsh, E.J. Knowledge does not protect against illusory truth. *J. Exp. Psychol. Gen.* **2015**, *144*, 993–1002. <https://doi.org/10.1037/xge0000098>
73. Hasher, L.; Goldstein, D.; Toppino, T. Frequency and the conference of referential validity. *J. Verbal Learn. Verbal Behav.* **1977**, *16*, 107–112. [https://doi.org/10.1016/S0022-5371\(77\)80012-1](https://doi.org/10.1016/S0022-5371(77)80012-1)
74. Bénabou, R.; Tirole, J. Mindful economics: The production, consumption, and value of beliefs. *J. Econ. Perspect.* **2016**, *30*, 141–164. <https://doi.org/10.1257/jep.30.3.141>
75. Kahan, D.M.; Peters, E.; Dawson, E.C.; Slovic, P. Motivated numeracy and enlightened self-government. *Behav. Public Policy* **2017**, *1*, 54–86. <https://doi.org/10.1017/bpp.2016.2>
76. Jost, J.T. *A Theory of System Justification*; Harvard University Press: Cambridge, MA, USA, **2020**.
77. Bandura, A. Moral disengagement in the perpetration of inhumanities. *Pers. Soc. Psychol. Rev.* **1999**, *3*, 193–209. https://doi.org/10.1207/s15327957pspr0303_3
78. Kruglanski, A.W.; Webster, D.M. Motivated closing of the mind: “Seizing” and “freezing.” *Psychol. Rev.* **1996**, *103*, 263–283. <https://doi.org/10.1037/0033-295X.103.2.263>
79. Glickman, M.; Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* **2025**, *9*, 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
80. Fanous, A.; Goldberg, J.; Agarwal, A.A.; Lin, J.; Zhou, A.; Daneshjou, R.; Koyejo, S. SycEval: Evaluating LLM sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, **2025**; Vol. 8, pp. 893–900.
81. Bo, J.Y.; Kazemitabaar, M.; Deng, M.; Inzlicht, M.; Anderson, A. Invisible saboteurs: Sycophantic LLMs mislead novices in problem-solving tasks. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2510.03667>
82. Kamenica, E.; Gentzkow, M. Bayesian persuasion. *Am. Econ. Rev.* **2011**, *101*, 2590–2615. <https://doi.org/10.1257/aer.101.6.2590>
83. Hill, K. They asked ChatGPT questions. The answers sent them spiraling. *The New York Times*, 13 June **2025**. Available online: <https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html> (accessed on 13 April 2026).
84. Dupré, M.H. People are becoming obsessed with ChatGPT and spiraling into severe delusions. *Futurism*, 28 June **2025**. Available online: <https://futurism.com/chatgpt-mental-health-crises> (accessed on 13 April 2026).
85. Gold, H. They thought they were making technological breakthroughs. It was an AI-sparked delusion. *CNN*, 30 June **2025**. Available online: <https://edition.cnn.com/2025/09/05/tech/ai-sparked-delusion-chatgpt> (accessed on 13 April 2026).
86. Schechner, S.; Kessler, S. ‘I feel like I’m going crazy’: ChatGPT fuels delusional spirals. *The Wall Street Journal*, 2 June **2025**. Available online: <https://www.wsj.com/tech/ai/i-feel-like-im-going-crazy-chatgpt-fuels-delusional-spirals-ae5a51fc> (accessed on 13 April 2026).
87. Hill, K. A teen was suicidal. ChatGPT was the friend he confided in. *The New York Times*, 26 August **2025**. Available online: <https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html> (accessed on 13 April 2026).
88. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
89. Dohnány, S.; Kurth-Nelson, Z.; Spens, E.; Luettgau, L.; Reid, A.; Gabriel, I.; Summerfield, C.; Nour, M.M. Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv* **2025**. <https://doi.org/10.48550/arXiv.2507.19218>

90. Edwards, W. Conservatism in human information processing. In *Formal Representation of Human Judgment*; Kleinmuntz, B., Ed.; Wiley: New York, NY, USA, **1968**; pp. 17–52.
91. Griffiths, T.L.; Kemp, C.; Tenenbaum, J.B. Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology*; Sun, R., Ed.; Cambridge University Press: Cambridge, UK, **2008**; pp. 59–100.
92. Payne, J.W.; Bettman, J.R.; Johnson, E.J. *The Adaptive Decision Maker*; Cambridge University Press: Cambridge, UK, **1993**.
93. von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, **1944**.
94. Thaler, R. Mental accounting and consumer choice. *Mark. Sci.* **1985**, *4*, 199–214. <https://doi.org/10.1287/mksc.4.3.199>
95. Kahneman, D.; Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **1979**, *47*, 263–291. <https://doi.org/10.2307/1914185>
96. Hochman, G.; Kalagy, T.; Malul, S.; Yosef, R. Choosing not to know: The emotional and sociocultural architecture of pension willful ignorance. *Curr. Opin. Psychol.* **2025**, 102181. <https://doi.org/10.1016/j.copsyc.2025.102181>
97. Munro, G.D.; Ditto, P.H. Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Pers. Soc. Psychol. Bull.* **1997**, *23*, 636–653. <https://doi.org/10.1177/0146167297236007>
98. Sunstein, C.R. *#Republic: Divided Democracy in the Age of Social Media*; Princeton University Press: Princeton, NJ, USA, **2017**.
99. YouGov. Americans **2024** Poll: AI Top Feeling Caution. Available online: <https://today.yougov.com/technology/articles/49099-americans-2024-poll-ai-top-feeling-caution> (accessed on 12 April 2026).
100. Mahmud, H.; Islam, A.N.; Ahmed, S.I.; Smolander, K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technol. Forecast. Soc. Change* **2022**, *175*, 121390. <https://doi.org/10.1016/j.techfore.2021.121390>
101. Gaczek, P.; Pozharliev, R.; Leszczyński, G.; Zieliński, M. Overcoming consumer resistance to AI in general health care. *J. Interact. Mark.* **2023**, *58*, 321–338. <https://doi.org/10.1177/10949968221151061>
102. Longoni, C.; Bonezzi, A.; Morewedge, C.K. Resistance to medical artificial intelligence. *J. Consum. Res.* **2019**, *46*, 629–650. <https://doi.org/10.1093/jcr/ucz013>
103. Dietvorst, B.J.; Simmons, J.P.; Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **2015**, *144*, 114. <https://psycnet.apa.org/doi/10.1037/xge0000033>
104. Beckers, J.J.; Schmidt, H.G. The structure of computer anxiety: A six-factor model. *Comput. Hum. Behav.* **2001**, *17*, 35–49. [https://doi.org/10.1016/S0747-5632\(00\)00036-4](https://doi.org/10.1016/S0747-5632(00)00036-4)
105. Brod, C. *Technostress: The Human Cost of the Computer Revolution*; Addison-Wesley: Reading, MA, USA, **1984**.
106. Meehl, P.E. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*; University of Minnesota Press: Minneapolis, MN, USA, **1954**.
107. Grove, W.M.; Zald, D.H.; Lebow, B.S.; Snitz, B.E.; Nelson, C. Clinical versus mechanical prediction: A meta-analysis. *Psychol. Assess.* **2000**, *12*, 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
108. Ayal, S.; Hochman, G. Ignorance or integration: The cognitive processes underlying choice behavior. *J. Behav. Decis. Mak.* **2009**, *22*, 455–474. <https://doi.org/10.1002/bdm.642>
109. Todd, P.M.; Gigerenzer, G. Environments that make us smart: Ecological rationality. *Curr. Dir. Psychol. Sci.* **2007**, *16*, 167–171. <https://doi.org/10.1111/j.1467-8721.2007.00497.x>
110. Barber, B.M.; Odean, T. Boys will be boys: Gender, overconfidence, and common stock investment. *Q. J. Econ.* **2001**, *116*, 261–292. <https://doi.org/10.1162/003355301556400>
111. Blanton, H.; Pelham, B.W.; DeHart, T.; Carvallo, M. Overconfidence as dissonance reduction. *J. Exp. Soc. Psychol.* **2001**, *37*, 373–385. <https://doi.org/10.1006/jesp.2000.1458>
112. Sperber, D.; Clément, F.; Heintz, C.; Mascaro, O.; Mercier, H.; Origg, G.; Wilson, D. Epistemic vigilance. *Mind Lang.* **2010**, *25*, 359–393. <https://doi.org/10.1111/j.1468-0017.2010.01394.x>

113. Steyvers, M.; Tejada, H.; Kumar, A.; Belem, C.; Karny, S.; Hu, X.; Mayer, L.W.; Smyth, P. What large language models know and what people think they know. *Nat. Mach. Intell.* **2025**, *7*, 221–231. <https://doi.org/10.1038/s42256-024-00976-7>
114. Hochman, G.; Erev, I. The partial-reinforcement extinction effect and the contingent-sampling hypothesis. *Psychon. Bull. Rev.* **2013**, *20*, 1336–1342. <https://doi.org/10.3758/s13423-013-0432-1>
115. Nass, C.; Moon, Y. Machines and mindlessness: Social responses to computers. *J. Soc. Issues* **2000**, *56*, 81–103. <https://doi.org/10.1111/0022-4537.00153>
116. Sundar, S.S. The MAIN model: A heuristic approach to understanding technology effects on credibility. In *Digital Media, Youth, and Credibility*; Metzger, M.J., Flanagin, A.J., Eds.; MIT Press: Cambridge, MA, USA, **2008**; pp. 73–100.
117. Metzger, M.J.; Flanagin, A.J. Credibility and trust of information in online environments: The use of cognitive heuristics. *J. Pragmat.* **2013**, *59*, 210–220. <https://doi.org/10.1016/j.pragma.2013.07.012>
118. Fernandes, D.; Villa, S.; Nicholls, S.; Haavisto, O.; Buschek, D.; Schmidt, A.; Kosch, T.; Shen, C.; Welsch, R. AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Comput. Hum. Behav.* **2026**, *175*, 108779. <https://doi.org/10.1016/j.chb.2025.108779>
119. Anderson, B.B.; Vance, A.; Kirwan, C.B.; Jenkins, J.L.; Eargle, D. From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *J. Manag. Inf. Syst.* **2016**, *33*, 713–743. <https://doi.org/10.1080/07421222.2016.1243947>
120. Bravo-Lillo, C.; Cranor, L.F.; Komanduri, S.; Schechter, S.; Sleeper, M. Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it. In *Proceedings of the Tenth Symposium on Usable Privacy and Security*, Menlo Park, CA, USA, 9–11 July 2014; USENIX Association: Berkeley, CA, USA, **2014**; pp. 105–111.
121. Obar, J.A.; Oeldorf-Hirsch, A. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Inf. Commun. Soc.* **2020**, *23*, 128–147. <https://doi.org/10.1080/1369118X.2018.1486870>
122. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. <https://doi.org/10.1145/3571730>
123. Duan, W.; Zhou, S.; Scalia, M.J.; Yin, X.; Weng, N.; Zhang, R.; Tolston, M. Understanding the evolvement of trust over time within human-AI teams. *Proc. ACM Hum.-Comput. Interact.* **2024**, *8* (CSCW2), 1–31. <https://doi.org/10.1145/3687060>
124. Fiske, S.T.; Taylor, S.E. *Social Cognition*; Addison-Wesley: Reading, MA, USA, **1984**.
125. Code, L. Epistemic responsibility. In *Virtue Epistemology*; Battaly, H., Ed.; Wiley-Blackwell: Chichester, UK, **2013**; pp. 109–126.
126. Kidd, I.J. Intellectual humility, confidence, and argumentation. *Topoi* **2016**, *35*, 395–402. <https://doi.org/10.1007/s11245-015-9324-5>
127. Krumrei-Mancuso, E.J.; Rouse, S.V. The development and validation of the Comprehensive Intellectual Humility Scale. *J. Pers. Assess.* **2016**, *98*, 209–221. <https://doi.org/10.1080/00223891.2015.1068174>
128. Tanesini, A. Intellectual humility as attitude. *Philos. Phenomenol. Res.* **2018**, *96*, 399–420. <https://doi.org/10.1111/phpr.12326>
129. Muyskens, K.; Ang, C.; Kerr, E.T. Relational epistemic humility in the clinical encounter. *J. Med. Ethics* **2025**, advance online publication. <https://doi.org/10.1136/jme-2024-110241>
130. Ivry, T. AI as moral cover: How algorithmic bias exploits psychological mechanisms to perpetuate social inequality. *Anal. Soc. Issues Public Policy* **2025**, *25*, e70031. <https://doi.org/10.1111/asap.70031>
131. Baer, M.; Kouchaki, M. Responsible collaboration with artificial intelligence in organizational scholarship: OBHDP's governance framework for authors and reviewers. *Organ. Behav. Hum. Decis. Process.* **2026**, *193*, 104480. <https://doi.org/10.1016/j.obhdp.2026.104480>

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.