

Cognitive Amplification vs Cognitive Delegation in Human–AI Systems: A Metric Framework

Eduardo Di Santi

March 20, 2026

Abstract

Artificial intelligence is increasingly embedded in human decision-making, where it can either enhance human reasoning or induce excessive cognitive dependence. This paper introduces a conceptual and mathematical framework for distinguishing *cognitive amplification*, in which AI improves hybrid human–AI performance while preserving human expertise, from *cognitive delegation*, in which reasoning is progressively outsourced to AI systems.

To characterize these regimes, we define a set of operational metrics: the Cognitive Amplification Index (CAI^*), the Dependency Ratio (D), the Human Reliance Index (HRI), and the Human Cognitive Drift Rate ($HCDR$). Together, these quantities provide a low-dimensional metric space for evaluating not only whether human–AI systems achieve genuine synergistic performance, but also whether such performance is cognitively sustainable for the human component over time.

The framework highlights a central design tension in human–AI systems: maximizing short-term hybrid capability does not necessarily preserve long-term human cognitive competence. We therefore argue that human–AI systems should be designed under a *cognitive sustainability* constraint, such that gains in hybrid performance do not come at the cost of degradation in human expertise.

1 Introduction

Artificial intelligence is rapidly becoming a central component of human decision-making across domains such as medicine, engineering, finance, and scientific research. In principle, AI systems can act as powerful amplifiers of human cognition, expanding the range of problems that individuals and teams can solve.

However, the increasing reliance on AI introduces a structural risk: humans may gradually delegate cognitive processes to automated systems, resulting in the erosion of analytical skills, domain understanding, and critical reasoning.

This paper proposes a simple but operational framework to analyze this phenomenon. We distinguish two regimes:

- **Cognitive Amplification:** AI increases the effective intelligence of human–AI systems.
- **Cognitive Delegation:** humans offload reasoning to AI systems, reducing their own cognitive capacity over time.

To formalize this distinction, we introduce a set of measurable quantities that describe the behavior of hybrid human–AI systems.

2 Related Work

The distinction between human cognition and artificial computation is a cornerstone of modern philosophy of mind and cognitive neuroscience. Central to this debate is the question of whether machines can replicate genuine understanding or if they merely simulate it. The Chinese Room argument [19] famously posits that syntactic manipulation of symbols, no matter how sophisticated, does not constitute semantic understanding. This foundational critique suggests that human–AI systems are not merely a union of two identical types of "intelligence," but rather a hybrid of different ontological processes.

The conceptual motivation for this paper is deeply rooted in the debates introduced by Byrne in his analysis of computation and consciousness [4]. By exploring the limits of functionalism and the Turing Test, Byrne highlights that the "mental" cannot be easily reduced to mere output, raising critical questions about what happens when humans begin to treat AI as a functional substitute for their own reasoning [5]. If the mind is "transparent" in its self-knowledge [3], then the delegation of cognitive steps to an opaque AI model represents a fundamental shift in how humans engage with their own belief-forming processes.

Furthermore, the biological grounding of human cognition suggests that intelligence is not a "dry" computational process. Solms argues that consciousness and the source of intentionality are rooted in affective, homeostatic

mechanisms—the "hidden spring" of the sentient brain [22]. This neuropsychanalytic perspective implies that human cognition is intrinsically tied to subjective experience and the Free Energy Principle [23], which prioritizes the reduction of uncertainty through active engagement with the environment. When humans delegate reasoning to AI, they risk bypassing the very "active loops" that characterize biological intelligence, potentially leading to the cognitive drift modeled in this work.

This risk is partially addressed by the extended mind hypothesis, which suggests that cognitive processes can and do extend into external artifacts [7]. However, while tools can expand our reach [20], empirical research on cognitive offloading shows that the tendency to minimize internal effort often leads to long-term changes in memory and reasoning strategies [18]. Recent studies in human–AI collaboration confirm that high-performance metrics in hybrid systems can mask a dangerous overreliance [25], where "automation bias" [14] and the inheritance of AI-driven errors [27] result in a net loss of human analytical competence.

Recent work on human–AI teaming emphasizes that the effectiveness of hybrid systems depends not only on AI capability but also on the quality of human–AI interaction and complementarity [1, 13, 15].

This paper proposes a conceptual and operational framework for analyzing human–AI collaboration. We distinguish two regimes: *cognitive amplification*, in which AI increases the effective problem-solving capacity of human–AI systems while preserving human expertise, and *cognitive delegation*, in which reasoning tasks are progressively outsourced to AI systems, leading to potential erosion of human cognitive capability.

To formalize this distinction, we introduce a set of measurable quantities that characterize the behavior of hybrid human–AI systems, including the Cognitive Amplification Index, the Dependency Ratio, and the Human Cognitive Drift Rate. Together, these metrics define a space of collaboration regimes that allows us to evaluate not only the performance of human–AI systems, but also their long-term cognitive sustainability.

3 System Model

Consider a hybrid system composed of a human and an AI agent.

$$S = H + A \tag{1}$$

where

- H represents human intelligence
- A represents artificial intelligence

We define the effective problem-solving capacity of the system as

$$Q(S) = Q(H, A) \tag{2}$$

where Q denotes the effective capability to solve tasks within a domain. Human intelligence alone corresponds to

$$Q_H = Q(H) \tag{3}$$

while AI capability alone corresponds to

$$Q_A = Q(A) \tag{4}$$

The hybrid system capability is therefore

$$Q_{HA} = Q(S) \tag{5}$$

4 Cognitive Amplification

In the ideal scenario, AI acts as a cognitive amplifier. The hybrid system exhibits synergy between human reasoning and machine exploration.

This can be modeled as

$$Q_{HA} = Q_H + Q_A + \alpha Q_H Q_A \tag{6}$$

where α represents the strength of human–AI interaction. If $\alpha > 0$, the system exhibits super-linear performance gains relative to the sum of the isolated components.[26, 25]

This expression should be interpreted as an idealized conceptual model of human–AI synergy, rather than as a directly observable empirical law. In practical applications, task-level performance measures such as accuracy, recall, F1 score, or time-to-solution are typically bounded, and empirical evaluation is therefore better conducted using relative performance metrics defined with respect to the best standalone agent.[26]

5 Cognitive Amplification Index

The original model $Q_{HA} = Q_H + Q_A + \alpha Q_H Q_A$ describes an idealized interaction between human and artificial intelligence. For empirical evaluation, however, it is more informative to measure how much the hybrid system improves over the best individual component.

We therefore define the Cognitive Amplification Index as

$$CAI^* = \frac{Q_{HA} - \max(Q_H, Q_A)}{\max(Q_H, Q_A)}. \quad (7)$$

This index directly quantifies the relative performance gain of the hybrid system compared to the best standalone agent.

CAI^*	Interpretation
> 0	Cognitive amplification over best component
$= 0$	Hybrid matches best component (no net gain)
< 0	Cognitive degradation (integration harms performance)

Values above zero indicate that human–AI collaboration produces a genuine synergistic effect, rather than merely reproducing the performance of either humans or AI alone.[26, 25]

6 Dependency Ratio

We next quantify how strongly hybrid performance depends on the AI component. Given a task-specific performance measure Q , we define the Dependency Ratio as

$$D = \frac{Q_A}{Q_{HA}}. \quad (8)$$

This ratio compares standalone AI performance to hybrid system performance and therefore serves as an operational measure of relative AI reliance within the hybrid configuration.

High values of D indicate that hybrid performance is close to the AI baseline, suggesting that the marginal contribution of the human component is limited and that the system may be operating in an AI-dominated regime. This interpretation is consistent with empirical work on overreliance and automation bias in human–AI teams.[14, 25]

For interpretive convenience, we also define a *Human Reliance Index* (HRI) as the complement of D :

$$HRI = 1 - D = \frac{Q_{HA} - Q_A}{Q_{HA}}. \quad (9)$$

Unlike D , the HRI does not provide independent information; rather, it expresses the same relation from the perspective of the human contribution to hybrid performance.

Range D	Range HRI	Qualitative regime
< 0.5	> 0.5	Human-dominant cognition
$0.5-0.8$	$0.2-0.5$	Balanced collaboration
> 0.8	< 0.2	AI-dominated cognition, risk of delegation

These thresholds are not universal constants, but operational regions that help distinguish human-dominant, balanced, and AI-dominated modes of collaboration in empirical studies of human–AI interaction.[26, 25]

7 Human Cognitive Drift

Recent empirical studies suggest that AI assistance can improve immediate task performance while reducing unassisted performance in subsequent evaluations [8].

To model long-term effects on human expertise, we consider the evolution of human-only performance over time. Let $Q_H(t)$ denote the problem-solving capability of the human when operating *without* AI assistance, evaluated through periodic “AI-off” assessment blocks in the same task domain.

We define the Human Cognitive Drift rate as

$$HCDR = \frac{Q_H(t_2) - Q_H(t_1)}{t_2 - t_1}. \quad (10)$$

Two regimes can occur:

- **Amplification regime**

$$HCDR \geq 0 \quad (11)$$

Human cognition is preserved or improved over time, even in the presence of AI assistance. This regime is more likely when AI tools scaffold human reasoning rather than replace it directly.[6, 25]

- **Delegation regime**

$$HCDR < 0 \tag{12}$$

Human cognition deteriorates over time due to excessive reliance on AI and reduced metacognitive monitoring. This phenomenon is closely related to automation bias, in which human operators tend to over-trust automated decision aids and reduce independent verification [21]. Empirical work in educational and decision-making settings shows that unrestricted access to AI can improve immediate performance while degrading unassisted performance when the AI is removed, consistent with negative human cognitive drift. Empirical research on cognitive offloading and human reliance on AI suggests that AI can improve immediate assisted performance while still undermining the user’s capacity to independently detect errors, critique outputs, or solve comparable tasks without assistance.[6, 24, 27]

In practical settings, $Q_H(t)$ and Q_{HA} can be approximated through task performance metrics such as problem-solving accuracy, decision quality, or task completion efficiency within a given operational domain.[6]

Figure 1 summarizes the two regimes of human–AI interaction considered in this work: cognitive amplification, in which AI enhances human reasoning while preserving human cognitive capacity, and cognitive delegation, in which excessive reliance on AI leads to progressive erosion of human expertise.

8 Regimes of Human–AI Collaboration

The metrics introduced above can be interpreted as defining a state space for human–AI systems. In particular, the pair (D, CAI^*) provides a useful low-dimensional representation of the interaction regime, where CAI^* captures the performance gain of the hybrid system over the best stand-alone component, and D measures the degree of AI dominance within the collaboration.

Figure 1 shows ...

Within this space, qualitatively distinct regimes can be identified. Human Cognitive Drift ($HCDR$) then determines whether a regime is stable from the standpoint of long-term human expertise or whether it tends toward progressive cognitive erosion.

Figure 2 shows ... These regimes illustrate that maximizing hybrid performance Q_{HA} does not necessarily guarantee cognitively sustainable human–AI collaboration.

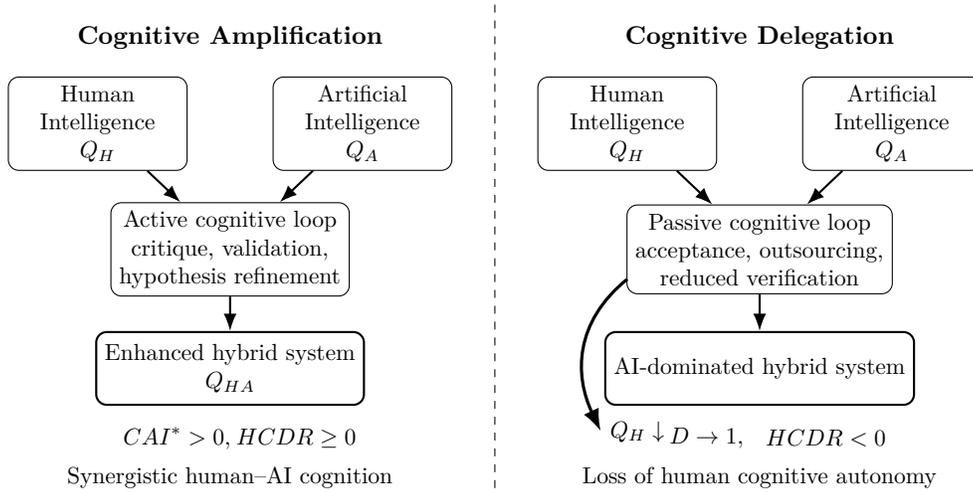


Figure 1: Two regimes of human–AI interaction. Left: cognitive amplification emerges when AI supports an active human cognitive loop. Right: cognitive delegation arises when reasoning is progressively outsourced to AI, increasing dependency and reducing human cognitive engagement over time.

9 Design Implications

The metric framework introduced above suggests concrete targets for the design of human–AI systems. In safety-critical domains, the goal is not only to maximize short-term hybrid performance, but to maintain *cognitive amplification* ($CAI^* > 0$), avoid excessive AI dominance (moderate D and non-trivial HRI), and keep human cognitive drift non-negative ($HCDR \geq 0$).

This leads to three high-level design objectives:

- sustain positive amplification at the system level ($CAI^* > 0$) rather than merely matching the best standalone component;
- avoid AI-dominated regimes in which $D \rightarrow 1$ and $HRI \rightarrow 0$, which are associated with overreliance and automation bias;
- preserve or improve human-only performance over time ($HCDR \geq 0$), preventing cognitive atrophy when AI assistance is removed.[6, 10]

These objectives can be translated into design principles for interfaces and workflows. Rather than providing final answers that humans merely

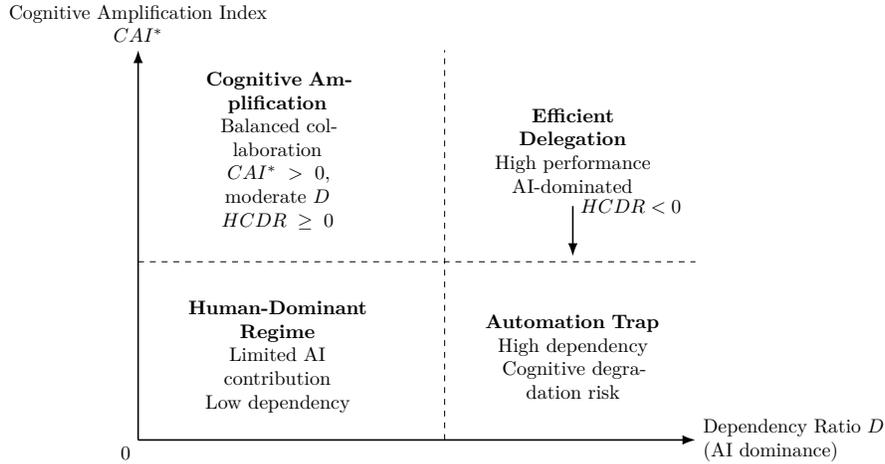


Figure 2: Conceptual phase diagram of human–AI collaboration regimes. The horizontal axis represents the Dependency Ratio D , measuring the degree of AI dominance in the hybrid system. The vertical axis represents the Cognitive Amplification Index CAI^* , indicating whether the human–AI system achieves genuine synergy over the best standalone component. Human Cognitive Drift ($HCDR$) determines whether a regime preserves or degrades long-term human expertise.

accept, AI systems should scaffold human reasoning and keep users in an *active* cognitive loop.

Possible design principles include:

- **Forcing explanation of reasoning steps.** Interfaces can require users to articulate their own hypothesis or rationale before revealing AI suggestions or detailed explanations. This reduces the risk that users simply adopt AI outputs without independent analysis and helps maintain $HCDR \geq 0$. [6, 12]
- **Exposing uncertainty and model confidence.** Communicating calibrated uncertainty, confidence intervals, or alternative candidate solutions helps users calibrate trust and reduces blind acceptance of AI recommendations. This is consistent with recent work on human–AI collaborative uncertainty quantification. [17, 9]
- **Requiring human hypothesis generation.** Interaction protocols can be structured so that the human must generate an initial diagnosis

or plan, and the AI acts as a critic or augments rather than as a primary oracle. Such designs increase the interpretive visibility of the human contribution and make it less likely that D approaches one.[10, 16]

- **Supporting exploration rather than final answers.** Systems can present evidence, counterfactual scenarios, or alternative solution paths that invite exploration, instead of single authoritative outputs. This shifts the role of AI from answer provider to exploration companion, encouraging deeper engagement and reducing the risk of long-term skill degradation.[10, 11]
- **Embedding AI-off evaluation blocks.** Periodic tasks in which users must operate without AI support enable direct measurement of $Q_H(t)$ and thus $HCDR$. These assessments can be integrated into training, simulation, or certification pipelines to detect early signs of cognitive delegation.[24, 6]

From an engineering perspective, these principles suggest that human–AI systems should be instrumented not only with standard performance metrics, but also with telemetry that allows continuous estimation of CAI^* , D , and $HCDR$. Such instrumentation would enable organizations to detect when their human–AI workflows drift from cognitive amplification toward cognitive delegation, and to redesign interfaces and training protocols accordingly.[16, 2]

10 Illustrative Measurement of Q

The framework introduced in this paper does not require a universal measure of intelligence. Instead, Q represents the effective problem-solving capability of a system within a specific task domain.

In practical settings, Q can be approximated using task-level performance metrics such as decision accuracy, solution quality, task completion time, or error rates in controlled problem-solving environments.

To illustrate this idea, consider a simplified engineering diagnostic task. An engineer must identify the root cause of a system anomaly from a set of sensor signals.

Suppose the following performance levels are observed:

System	Diagnostic Accuracy
Human alone (Q_H)	0.70
AI alone (Q_A)	0.80
Human + AI (Q_{HA})	0.92

The Cognitive Amplification Index is then estimated as

$$CAI^* = \frac{Q_{HA} - \max(Q_H, Q_A)}{\max(Q_H, Q_A)}. \quad (13)$$

Substituting the observed values:

$$CAI^* = \frac{0.92 - 0.80}{0.80} \approx 0.15. \quad (14)$$

The Dependency Ratio becomes

$$D = \frac{Q_A}{Q_{HA}} = \frac{0.80}{0.92} \approx 0.87. \quad (15)$$

Its complementary Human Reliance Index is therefore

$$HRI = 1 - D \approx 0.13. \quad (16)$$

In this case the hybrid system exhibits a clear gain over the best individual component ($CAI^* > 0$). However, the high value of D indicates that hybrid performance remains strongly anchored to the AI baseline, suggesting an AI-dominated configuration despite the positive value of CAI^* . This pattern is consistent with empirical findings on overreliance, where human–AI teams can achieve high immediate performance while risking long-term degradation of human expertise.[25, 27]

10.1 Illustrative Examples of the Framework

Pharmaceutical Safety Monitoring. Consider a pharmacovigilance task in which safety analysts must identify potential adverse drug reactions from large collections of clinical reports.

Let Q represent the ability to correctly identify safety signals. Performance can be measured through standard metrics such as recall, precision, or F1 score.

Suppose the following results are observed in a controlled evaluation:

System	Signal Detection Recall
Human experts (Q_H)	0.72
AI model (Q_A)	0.78
Human + AI collaboration (Q_{HA})	0.91

The resulting Cognitive Amplification Index is

$$CAI^* = \frac{0.91 - 0.78}{0.78} \approx 0.17.$$

The Dependency Ratio becomes

$$D = \frac{0.78}{0.91} \approx 0.86.$$

The complementary Human Reliance Index is

$$HRI = 1 - D \approx 0.14.$$

In this scenario, the hybrid system clearly outperforms both humans and the AI model alone ($CAI^* > 0$), indicating cognitive amplification at the system level. However, the high value of D suggests that hybrid performance remains close to the AI baseline, placing the team in an AI-dominated regime. If analysts primarily accept model suggestions without actively interrogating them, this configuration may drift toward cognitive delegation, with a risk of long-term degradation of $Q_H(t)$. [24, 27]

Industrial Anomaly Diagnosis. Consider an engineering monitoring system in which operators must diagnose anomalies in a complex industrial process using sensor data.

Let Q represent diagnostic accuracy under time constraints. Performance can be measured as the proportion of correctly identified failure modes during simulated operational scenarios.

Assume the following experimental results:

System	Diagnostic Accuracy
Human operator (Q_H)	0.68
AI diagnostic system (Q_A)	0.81
Human + AI collaboration (Q_{HA})	0.95

The Cognitive Amplification Index becomes

$$CAI^* = \frac{0.95 - 0.81}{0.81} \approx 0.17.$$

The corresponding Dependency Ratio is

$$D = \frac{0.81}{0.95} \approx 0.85.$$

The complementary Human Reliance Index is

$$HRI = 1 - D \approx 0.15.$$

If the hybrid interface is designed so that operators must generate, critique, and refine diagnostic hypotheses—for example, by requiring explanation of reasoning steps and uncertainty inspection—the system may sustain a regime of cognitive amplification with non-negative human cognitive drift. By contrast, if operators routinely outsource diagnosis to the AI, empirical findings on automation bias and cognitive offloading suggest that $Q_H(t)$ may deteriorate, pushing the system into a delegation regime despite high short-term performance.[14, 6, 24]

11 Discussion

The framework proposed here does not attempt to measure intelligence in an absolute sense. Instead, it provides a relative metric system to evaluate whether AI systems amplify or replace human cognition.

As AI systems become more capable, this distinction becomes critical for maintaining human intellectual autonomy.

Safety-critical industries such as railway systems, aviation, pharmaceutical manufacturing, and medical technology depend on highly trained human expertise for decision making under uncertainty. In these domains, artificial intelligence should not replace human judgment but rather amplify the cognitive capabilities of expert operators. This challenge is closely related to well-documented phenomena such as automation bias, in which operators may over-trust automated decision aids and reduce independent verification [14, 25].

This observation highlights a fundamental tension in the design of human–AI systems: maximizing short-term hybrid performance does not necessarily preserve long-term human expertise. A system that maximizes Q_{HA} while driving $HCDR < 0$ may achieve short-term efficiency at the cost of a progressive loss of human expertise.

In this sense, the design of human–AI systems should be guided not only by immediate hybrid performance, but also by the long-term *cognitive sustainability* of human expertise. The framework proposed in this paper provides a way to evaluate whether human–AI systems operate in a regime of cognitive amplification or cognitive delegation, and whether such regimes are sustainable over time.

This distinction is particularly relevant in safety-critical environments, where the degradation of human expertise due to excessive reliance on automation may introduce systemic risks [25, 14].

The present framework therefore suggests that human–AI systems should be optimized under a sustainability constraint: improvements in hybrid performance should not come at the expense of negative human cognitive drift. Formally, this principle can be expressed as

$$\max Q_{HA} \quad \text{subject to} \quad HCDR \geq 0,$$

which states that gains in hybrid system capability should not be achieved by degrading the long-term cognitive competence of the human component.

12 Conclusion

Artificial intelligence presents a dual possibility: it can amplify human intelligence or gradually replace it.

The metric framework proposed in this paper provides a way to analyze and monitor this transition. Future work should focus on empirical evaluation of these metrics in real-world human–AI interaction settings.

In safety-critical domains, the goal of artificial intelligence should not be the replacement of human cognition but the amplification of expert judgment. Systems that increase automation while degrading human expertise may ultimately reduce, rather than increase, the intelligence of the overall system.

References

- [1] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the AAI Conference on Artificial Intelligence*, 2021.

- [2] Alexander Borg and Sandra Wachter. Automation bias in the EU AI act: On the legal implications of human oversight. *European Journal of Risk Regulation*, 16(1):1–24, 2025.
- [3] Alex Byrne. *Transparency and Self-Knowledge*. Oxford University Press, 2018.
- [4] Alex Byrne. Minds and machines. MITx Online / MIT Open Learning Library, 2019.
- [5] Alex Byrne and Jaegwon Kim. *Philosophy of Mind*. Westview Press, 2012.
- [6] Gita Chirayath and Daniel Gerlich. Cognitive offloading or cognitive overload? how AI alters the mental architecture of coping. *Frontiers in Psychology*, 16, 2025.
- [7] Andy Clark and David J. Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.
- [8] Fabrizio Dell’Acqua, Edward McFowland, and Ethan Mollick. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Working Paper*, 2023.
- [9] Alan Dix and colleagues. Uncertainty, explainability, transparency, and bias in AI. Northumbria University, 2020.
- [10] Daniel Gerlich. Designing AI for human expertise: Preventing cognitive shortcuts. *UXmatters*, February 2025.
- [11] Daniel Gerlich. AI’s cognitive implications: The decline of our thinking skills? *IE Insights*, 2026.
- [12] Judy W. Gichoya et al. AI pitfalls and what not to do: Mitigating bias in AI. *npj Digital Medicine*, 6:136, 2023.
- [13] Matthew Gombolay, Rikke Jensen, and Julie Shah. Human–ai teaming: Foundations and challenges. *ACM Computing Surveys*, 2023.
- [14] Jessica Green et al. Bending the automation bias curve: A study of human- and AI-based decision making. *International Studies Quarterly*, 68(2), 2024.

- [15] Ece Kamar. Complementing ai systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 4070–4073, 2016.
- [16] Thomas W. Malone. Designing the intelligent organization. *MIT Sloan Management Review*, 62(3), 2021.
- [17] Shayan Noorani et al. Human–AI collaborative uncertainty quantification. *arXiv preprint*, 2025.
- [18] Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.
- [19] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.
- [20] Herbert A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 3 edition, 1996.
- [21] Linda Skitka, Kathleen Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 1999.
- [22] Mark Solms. *The Hidden Spring: A Journey to the Source of Consciousness*. W. W. Norton, New York, 2021.
- [23] Mark Solms and Karl Friston. The hard problem of consciousness and the free energy principle. *Frontiers in Psychology*, 9:2714, 2019.
- [24] Aigul Sultanova, Michael Evans, and Ji-Hyun Park. The impact of artificial intelligence tools on human cognitive abilities. *Innovation: Technology, Governance, Globalization*, 6, 2025.
- [25] Kristen Vaccaro, Jim Waldo, et al. Overreliance on AI: Literature review. Technical report, Microsoft Aether Working Group, June 2022.
- [26] Michelle Vaccaro, Abdullah Almaatouq, and Thomas W. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303, 2024.
- [27] Lucía Vicente and Helena Matute. Humans inherit artificial intelligence biases. *Scientific Reports*, 13:15737, 2023.