Human Factors and Systems Interaction, Vol. 230, 2026, 27–38
https://doi.org/10.54941/ahfe1007974

From Task to Intentionality Automation: Mitigating the Open-Loop and Metacognitive Gaps in Agentic AI

Mário Simões-Marques

CINAV, Escola Naval, Instituto Universitário Militar, Base Naval de Lisboa, 2810–001 Almada, Portugal

ABSTRACT

Artificial Intelligence (AI) enables powerful capabilities that are transforming almost all sectors. However, the economic growth driven by AI comes at a cost, and its sociotechnical impacts are fraught with contradictions and paradoxes. As a result, several legal initiatives and risk management frameworks have been introduced to mitigate the various risks associated with AI systems. Agentic AI systems require even closer attention than traditional AI. While traditional AI has a narrow focus and responds to direct commands, Agentic AI emerges from combining multiple types of AI capable of planning, tool use, and multi-step execution. These systems can behave and interact autonomously, making decisions and performing tasks to achieve system objectives with minimal human oversight. Recognizing that Agentic AI represents a paradigm shift, this paper addresses its challenges from a Human-AI Interaction perspective. It examines the root causes and impacts of risks arising from the transition from Task Automation to Intentionality Automation, where the user manages outcomes and constraints rather than individual task steps. Key issues include the Open-Loop Control Gap and the Metacognitive Gap, whose relationship is fundamental to understanding the collapse of human oversight, as they represent two sides of the same coin in the loss of control. By analysing scenarios such as cybersecurity and healthcare, this paper identifies dimensions of user demand and identifies Ecological Interface Design as an ergonomic approach to ensure that as AI gains agency, the human retains authority and situational awareness.

Keywords: Agentic AI, Human factors, Intentionality automation, Metacognitive gap, Human-AI interaction, Human oversight

INTRODUCTION

The rapid advancement of agentic artificial intelligence (AI) systems—autonomous agents powered by large language models (LLMs) with capabilities for planning, tool use, memory, and self-directed action—represents a fundamental shift in human-AI interaction paradigms. Unlike traditional AI systems that operate as reactive tools responding to explicit user commands, agentic AI systems possess the capacity to independently pursue goals, make sequential decisions, and execute complex multi-step tasks with minimal human oversight (Chhabra et al., 2026). This evolution promises unprecedented productivity gains across domains ranging from

Received February 27, 2026; Revised April 2, 2026; Accepted April 16, 2026; Available online July 20, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

software development and scientific research to healthcare and enterprise automation (Ransbotham et al., 2025). However, this transition from tool-based to agent-based AI introduces profound challenges related to how humans delegate intentionality, maintain oversight, and preserve meaningful control over increasingly autonomous systems.

As noted in (Rosenbacke et al., 2025), today's LLMs don't actually apply reasoning to generate their answers—they predict text statistically. By analysing patterns in massive amounts of training data, they estimate which word is most likely to follow the previous one, producing responses that sound natural and fluent. However, this mechanism has a fundamental weakness: these models have no genuine grasp of meaning, and they cannot independently verify whether what they're saying is true or accurate. This means users can receive answers that feel convincing and well-structured while being factually wrong, logically inconsistent, or entirely made up. In the AI field, this problem is known as "hallucination" — when a model generates information that has no real basis, or draws incorrect conclusions from the data it was given. For writing this paper, the author interacted with different "traditional" LLM platforms in the exploratory phase and experienced this problem. Despite receiving very compelling answers, many of them did not survive the reliability check. The LLM models invented plausible-sounding author names and titles, misattributed real concepts to wrong papers, cited papers that don't actually exist, and provided wrong publication years/venues.

The transition from Generative AI to Agentic AI represents a fundamental shift in the human-machine hierarchy. While traditional LLMs function as passive tools for content synthesis (corresponding to Task Automation), agentic systems engage in Intentionality Automation, characterized by autonomous goal-decomposition, iterative planning, and the execution of tasks across digital environments (Acharya et al., 2025). In this paradigm, artificial agents do not merely suggest text or images; they interpret high-level human objectives, decompose them into sub-tasks, and execute them autonomously. This delegation of "intent" introduces significant systemic risks, which are potentially much more severe, compared to the use of traditional AI, since the level of human supervision tends to be reduced, and an overreliance on AI potentiates engaging these systems in the automation of critical activities, such as healthcare or cybersecurity.

Central to these challenges are two interrelated gaps, here identified as Open-Loop Gap and Metacognitive Gap, that emerge when intentionality the capacity to direct actions toward specific goals—becomes automated within AI agents.

This paper addresses these gaps by proposing mitigation strategies specifically designed to bridge the Open-Loop and Metacognitive Gaps in agentic AI systems. By synthesizing insights from human-computer interaction, cognitive science, AI security, and healthcare studies, this paper aims to establish foundational principles for designing the interaction with agentic systems that enhance rather than undermine human cognitive capabilities and autonomous decision-making.

AGENTIC SYSTEMS RELATED GAPS

The Open-Loop Gap

The Open-Loop Gap refers to the absence of continuous feedback mechanisms that would allow users to maintain real-time awareness of an agent's state, reasoning process, and ongoing actions. As Holzinger et al. (2025) observe, the complexity of contemporary AI architectures, particularly large-scale neural networks and generative AI applications, fundamentally undermines human understanding and decision-making capabilities, leading them to conclude that complete oversight may no longer be viable in certain contexts. This observation is particularly salient in agentic systems where agents operate across extended time horizons, potentially making hundreds of decisions before presenting results to users. In fact, the agents can execute multi-step workflows without intermediate human validation, potentially leading to "intent drift" (Bandi et al., 2025). The European Data Protection Supervisor's analysis further emphasizes that human oversight alone cannot serve as the ultimate solution for ensuring the safety of partially or fully automated decision-making systems, highlighting the inadequacy of traditional supervisory approaches when applied to autonomous agents (EDPS, 2025).

The Metacognitive Gap

The Metacognitive Gap describes the disconnect between users' perceived understanding and control over AI-assisted processes and their actual cognitive capacity to monitor, evaluate, and regulate these interactions. On other words, the Metacognitive Gap describes the misalignment between the agent's internal uncertainty and the human's perceived trust, often exacerbated by the system's "surface fluency" — the ability to appear competent without a corresponding depth of reasoning (Atf & Lewis, 2025; Zhang et al., 2025). This results both from the failure of the system to communicate its internal uncertainty and from the failure of the human to accurately calibrate trust, often resulting in over-reliance or automation bias (Li et al., 2025).

AGENTIC SYSTEMS IMPACTS IN COGNITIVE LOAD

Recent research indicates that as agents become more autonomous, the human role shifts from "executor" to "supervisor," yet current interfaces are ill-equipped for this transition. Therefore, the user's mental resources are redirected from task execution to intent monitoring. This shift can inadvertently increase Extraneous Load (workload caused by poorly designed interfaces) while depleting the Germane Load (resources used for high-level reasoning and supervision). Thus, the cognitive load does not decrease; rather, it transforms into a high-level monitoring burden that humans are biologically unsuited for over long durations (Shen et al., 2025). This finding is also termed in recent research as the "AI Placebo Effect" (Kosch et al., 2022). When users believe they are being supported by a sophisticated agent, their subjective confidence increases, but their metacognitive accuracy—the ability to judge their own performance or the agent's errors—decreases. This creates a hidden cognitive load: the "monitoring cost" of verifying an agent that appears highly competent but lacks situational awareness (Sidra & Mason, 2025).

Current studies highlight a critical lack of "traceability" in agentic reasoning, where the "why" behind an agent's specific tool-use or sub-goal is obscured. This opacity makes it nearly impossible for users to intervene before a catastrophic error occurs, particularly in high-stakes domains such as cybersecurity and clinical healthcare (Lazer et al., 2026; Shin, 2025).

Tankelevitch and colleagues provide a comprehensive framework for analysing the metacognitive demands of generative AI, identifying critical components including self-awareness, confidence calibration, metacognitive flexibility, and task decomposition abilities (Tankelevitch et al., 2024). Their research reveals that while automated suggestions reflect tighter integration between manual work and AI, reducing some metacognitive demands associated with explicit prompting, they nevertheless introduce challenges such as cognitive interruptions and diminished reflective awareness. This finding is corroborated by Fernandes and colleagues, who demonstrated a troubling performance-metacognition disconnect: while AI assistance improved task performance on the reasoning tasks, it simultaneously led to substantial overestimation of users' own capabilities, indicating severely compromised metacognitive accuracy (Fernandes et al., 2025). Paradoxically, their study found that increased technological knowledge and critical appraisal of AI, as measured by standardized literacy scales, actually increased user confidence while simultaneously decreasing the accuracy of self-assessment—a phenomenon suggesting that familiarity with AI systems may produce false confidence rather than genuine metacognitive insight.

The implications of the above mentioned gaps are amplified by the well-documented phenomenon of Automation Bias, as the tendency for human operators to over-rely on automation such that automated systems and their outputs become a heuristic replacement for vigilant information seeking and processing (Horowitz & Kahn, 2024). A comprehensive review of automation bias in human-AI collaboration done by Romeo and Conti reveals that complex explanations designed to support user understanding paradoxically increased cognitive load, hindering effective processing, while increased workload induced by decision difficulty contributed to overreliance by adding cognitive burden (Romeo & Conti, 2026). This creates a vicious cycle: as agentic systems become more capable and autonomous, they impose greater metacognitive demands on users, yet simultaneously their design provide fewer opportunities for users to develop the skills necessary to meet those demands. Kabashkin formalized this dynamic through a cognitive co-evolution model, introducing the Cognitive Sustainability Index as a composite measure integrating autonomy, reflection, creativity, delegation, and reliance (Kabashkin, 2025). This author demonstrated that when users interact with AI under time pressure or cognitive overload—conditions increasingly common in agentic workflows—reliance on automation intensifies, shifting the system toward what the author term "cognitive atrophy."

RISKS IN AGENTIC SYSTEMS

Beyond technical vulnerabilities, agentic systems introduce novel risks related to sense of agency and responsibility attribution. Research examining human-agent interactions through the lens of social agency reveals that when a system makes autonomous decisions, the sense of initiation and intentionality is most often automatically attributed to the system itself, potentially creating illusory control where users believe they maintain oversight that does not correspond to any objective control over ongoing operations (Strebinger & Treiblmaier, 2025).

Despite growing recognition of these challenges, significant research gaps remain. While substantial work has examined cognitive challenges in human-AI delegation for discrete classification tasks (Fügener et al., 2022), far less attention has been devoted to the longitudinal, multi-step decision processes characteristic of agentic workflows. Similarly, while explainable AI has been proposed as a mechanism to support appropriate reliance by enabling humans to follow AI predictions when accurate or override them when incorrect (Senoner et al., 2024), the effectiveness of XAI techniques for maintaining metacognitive awareness across extended agentic interactions remains largely unexplored. Measurement frameworks for agency in human-robot interaction have primarily employed either self-determination theory with psychometric measures or neuroscientific intentional binding paradigms (Glawe et al., 2025), yet these approaches have not been systematically adapted to the unique characteristics of LLM-based agentic systems where the locus of intentionality becomes fundamentally ambiguous.

Nevertheless, even a cursory review of the academic literature reveals that many other authors are highlighting the challenges and risks of Agentic AI from multiple perspectives (e.g., cognition/trust/reliance: (Buçinca et al., 2021; Klingbeil et al., 2024; Maynard, 2026; Mehrotra et al., 2024; Park et al., 2024; Sahebi & Formosa, 2025); cybersecurity: (Chhabra et al., 2026; Lazer et al., 2026; Sapkota et al., 2026); or healthcare: (Goktas & Grzybowski, 2025; Holzinger et al., 2025; Jung et al., 2025; Salehi et al., 2025; Tikhomirov et al., 2024; Verma et al., 2025)).

The risks associated with agentic AI are not merely theoretical; they represent structural vulnerabilities in how autonomous software interacts with the physical and digital world. The next subsections offer some insights on the risks identified in Cybersecurity and Healthcare agentic systems.

Evidence of Risks in Cybersecurity Agentic Systems

In Cybersecurity, agentic systems expand the attack surface by creating "non-human identities" when the agent is granted the authority to execute code or has over-privileged access to sensitive APIs (WEF, 2025).

Examples of this are provided in an academic survey done by Lazer and colleagues on Agentic AI and Cybersecurity which, for instance, highlights studies addressing the risk of Indirect Prompt Injection, where malicious code embedded in external data (e.g., a retrieved document or website)

hijacks the agent's "intentionality" to perform unauthorized actions, such as data exfiltration or credential theft (Lazer et al., 2026). Recent research has quantified the security vulnerabilities inherent in agentic architectures, reporting that 94.4% of state-of-the-art LLM agents are vulnerable to Direct Prompt Injection attacks, 83.3% to retrieval-based backdoors, and 100% to Inter-Agent Trust Exploitation attacks (Lupinacci et al., 2025).

The industry also warns on the emerging risks added by agentic systems, namely regarding Cascading Failures. For instance, the company McKinsey identifies "Chained Vulnerabilities," where a logic error in an initial agent (e.g., a data-processing agent) propagates through a multi-agent ecosystem, leading to high-magnitude failures that are difficult to trace back to the source (Klein et al., 2025).

In these examples the structural risk mainly relates with the Open-Loop Gap, where the agent may perceive the malicious instruction as a valid sub-task necessary to achieve the user's original goal, effectively hijacking the user's intentionality.

Evidence of Risks in Healthcare Agentic Systems

Recent research studies — e.g., (Griot et al., 2025; Khan et al., 2025; Räz et al., 2025; Yu et al., 2025) — support the assertion that trust calibration failures occur when clinicians rely on an AI agent's fluency rather than its clinical explainability. The studies analysed evidence several risks, namely:

- Trust Calibration Failures: AI users, including clinicians, often exhibit dual, simultaneous failures: automation bias (over-reliance on the agent) or algorithm aversion (excessive rejection after errors);
- Fluency Over Explanation: the conversational fluency of an agent can create a sense of trust that bypasses necessary epistemic vigilance (the critical evaluation of the output's accuracy);
- Affective vs. Cognitive Trust: when AI systems appear empathetic or articulate, they foster "affective trust" (rooted in comfort and perceived warmth). This often overshadows "cognitive trust," which should be based on clinical validation and rational assessment of performance;
- Limitations of Explainability: explainability tools are meant to aid trust; yet, clinicians are not always aware of their practical value, and sometimes these tools provide "visually plausible" but clinically inaccurate justifications;
- Expertise Paradox: occurs where novices over-rely on AI, while experts might under-rely on it, both stemming from faulty calibration.

In these examples the structural risk is mainly related to the Metacognitive Gap, which is a critical vulnerability in clinical AI, where the agent exhibits a disconnect between their confidence and their actual accuracy, leading clinicians to trust faulty, yet highly confident, advice. While LLMs can achieve high, expert-level performance on medical benchmarks, they often

fail to recognize the limits of their own knowledge, presenting incorrect or flawed reasoning with supreme linguistic certainty. Therefore, several authors suggest that for AI to be truly trustworthy in clinical settings, it must be designed with "epistemic vigilance" in mind—encouraging clinicians to verify, not just trust based on smooth, fluent, but potentially inaccurate, outputs (Khan et al., 2025).

Besides Metacognitive Gap related risks, research also addresses Clinical Safety risks related to the Open-Loop Gap, for instance where agentic AI is deployed to manage longitudinal care planning, if an agent autonomously modifies a patient's monitoring schedule based on an incorrect interpretation of lab results, and the system does not surface the reasoning for that change, the error may go undetected until a sentinel event occurs (Shin, 2025).

A MITIGATION PROPOSAL: ECOLOGICAL INTERFACE DESIGN

To bridge the gaps in intentionality automation, research propose a shift from navigation-centric UX toward Ecological Interface Design (EID). Originally presented in (Vicente & Rasmussen, 1992) and developed for high-risk environments like nuclear power plants, EID focuses on making the "constraints" and "affordances" of the work domain visible to the operator. Therefore, such approach can contribute to improve the Human-Agentic AI Interaction by making the underlying constraints of the system visible to the operator. Table 2 provides a comparison of the features of navigation-centric design vs. EID.

Bridging the Open-Loop Gap

This gap is mitigated through EID by replacing traditional "black-box" progress bars with Functional Abstraction Hierarchies (FAH). Instead of seeing that an agent is "working," the user sees a real-time map of the agent's goal decomposition.

By mapping the agent's operations from Functional Purpose (the "Why") down to Physical Form (the "How"), EID provides a continuous "intent-status" synchronization allowing the supervisor to maintain situational awareness without being overwhelmed by data points. The envisaged strategy is implementing "Semantic Checkpoints" where the agent must present its interpreted sub-goals for human validation before execution (Panchal, 2025). By forcing users' interaction through a Human-in-the-Loop approach, this solution allows users to identify intent drift early, effectively closing the Open-Loop Gap.

In summary, EID can mitigate the constraints of Miller's Law by using a Functional Abstraction Hierarchy (FAH). FAH is a critical EID tool for closing the Open-Loop Gap. It maps the agent's operations across four levels by "chunking" complex agentic data into digestible levels of intent. The description of the FAH levels and the cognitive function they serve are summarized in Table 1.

Table 1: Ecological interface design functional abstraction hierarchies levels.

FAH Level	Description	Cognitive Function
Functional Purpose	"Why": High-level human intent.	Maintains Goal Alignment.
Abstract Function	"Logic": Planning and causal models.	Closes the Metacognitive Gap.
Generalized Function	"Tools": Specific APIs and actions.	Respects via chunking.
Physical Form	"Execution": Data flows and code.	Provides Traceability.

Table 2: Traditional UX vs. ecological interface design (EID) features.

Feature	Traditional UX (Navigation-Based)	Ecological Interface Design (Intent-Based)
User Role	Direct actor (clicking, typing)	Strategic supervisor (orchestration)
System Status	Black-box "Processing..." bars	Functional Abstraction Hierarchy (FAH) real-time map
Goal Visibility	Only current step is visible	Mapping of sub-goals to high-level intent
Feedback Loop	Reactive (post-error alerts)	Proactive (visibility of constraints/drift)

Closing the Metacognitive Gap

Closing the Metacognitive Gap requires interfaces that facilitate Mutual Adaptation. The goal is to provide the user a window that offers Uncertainty Visualization. Rather than a single output, EID-based interfaces surface the "reasoning paths" not taken, allowing the human to see the agent's level of certainty across different branches of the task (Shen et al., 2025).

Another approach is based on the Socratic Interaction Model. This approach suggests that agents should prompt users with "clarification requests" when a task's intent falls into a low-confidence region of the agent's model, preventing the user from slipping into passive monitoring (Kamalov et al., 2026).

In summary, Intentionality Automation is not a binary state but a spectrum that must be managed through "Transparent Intentionality." Open-Loop Gap is primarily an information-flow problem. Traditional dashboards only report events (the loop is closed only at the end). EID closes the loop semantically by providing a continuous "Intent-Status" synchronization. When the agent moves from planning to execution, the FAH surfaces the transition, allowing the human to act as a "gatekeeper" of intentionality without needing to execute the task themselves (Bandi et al., 2025; Panchal, 2025).

The Metacognitive Gap is often widened by "black-box" agents that prioritize efficiency over explainability. Reported findings indicate that Uncertainty Visualization — a core component of EID — acts as a metacognitive prompt. By visualizing the "reasoning paths not taken," the interface forces the user into a state of Epistemic Vigilance (Maynard, 2026). This prevents the "AI Placebo Effect" where users stop critically evaluating the agent's outputs (Kosch et al., 2022).

CONCLUSION

The delegation of intent to AI agents is inevitable, but it does not have to be opaque. This paper discussed how Ecological Interface Design can contribute to mitigating the risks of Agentic AI Systems, namely in critical domains such as cybersecurity and healthcare by: i) Reducing the Extraneous Cognitive Load required to interpret agent behaviour; ii) Closing the Open-Loop Gap through real-time mapping of functional purpose; and iii) Enhancing Metacognitive Calibration to prevent dangerous over-reliance. The move toward Agentic AI necessitates an interaction philosophy that respects the limits of human cognition. By integrating Miller's Law into Ecological Interface Design, we can create systems where intentionality is automated but remains transparent. Closing the Open-Loop and Metacognitive Gaps ensures that as AI agents become more autonomous, they remain fundamentally aligned with human oversight and safety.

ACKNOWLEDGMENT

The work was funded by the Portuguese Navy.

REFERENCES

Acharya, D., Kuppan, K., & Ashwin, D. B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals – A Comprehensive Survey. *IEEE Access*, *PP*, 1–1.

Atf, Z., & Lewis, P. R. (2025). Is Trust Correlated With Explainability in AI? A Meta-Analysis. *IEEE Transactions on Technology and Society*, 1–8.

Bandi, A., Kongari, B., Naguru, R., Pasnoor, S., & Vilipala, S. V. (2025). The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet*, *17*(9).

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW1), 188:1-188:21.

Chhabra, A., Datta, S., Nahin, S. K., & Mohapatra, P. (2026). *Agentic AI Security: Threats, Defenses, Evaluation, and Open Challenges* (arXiv:2510.23883). arXiv.

EDPS. (2025). *Human Oversight of Automated Decision-Making | European Data Protection Supervisor (TechDispatch #2/2025)*.

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2025). Performance and Metacognition Disconnect when Reasoning in Human-AI Interaction. *Computers in Human Behavior*, *175*, 108779.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2), 678–696.

Glawe, F., Schmeckel, T., Brauner, P., & Ziefle, M. (2025). *Human Autonomy and Sense of Agency in Human-Robot Interaction: A Systematic Literature Review* (arXiv:2509.22271). arXiv.

Goktas, P., & Grzybowski, A. (2025). Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *J. of Clinical Medicine*, *14*(5), 1605.

Griot, M., Hemptinne, C., Vanderdonckt, J., & Yuksel, D. (2025). Large Language Models lack essential metacognition for reliable medical reasoning. *Nature Comm.*, *16*, 642.

Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, *85*, 59–62.

Horowitz, M. C., & Kahn, L. (2024). Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly*, *68*(2), sqae020.

Jung, J., Phillipi, M., Tran, B., Chen, K., Chan, N., Ho, E., Sun, S., & Houshyar, R. (2025). Accuracy of large language models in generating differential diagnosis from clinical presentation and imaging findings in pediatric cases. *Ped. Rad.*, *55*(9), 1927–1933.

Kabashkin, I. (2025). Cognitive Atrophy Paradox of AI–Human Interaction: From Cognitive Growth and Atrophy to Balance. *Information*, *16*(11).

Kamalov, F., Calonge, D. S., Smail, L., Azizov, D., Thadani, D. R., Kwong, T., & Atif, A. (2026). *Evolution of AI in Education: Agentic Workflows* (arXiv:2504.20082). arXiv.

Khan, M., Fong, C., & Tripathi, S. (2025). "Trust, but Verify": A Reflexive Thematic Analysis of Human–AI Interaction. *Advances in Social Sciences Research Journal*, *12*, 237–251. https://doi.org/10.14738/assrj.1211.19642

Klein, B., Lewis, C., & Isenberg, R. (2025). Deploying agentic AI with safety and security: A playbook for technology leaders. *McKinsey Quarterly*.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352.

Kosch, T., Welsch, R., Chuang, L., & Schmidt, A. (2022). The Placebo Effect of Artificial Intelligence in Human-Computer Interaction. *ACM Transactions on Computer-Human Interaction*, *29*(6), 1–32.

Lazer, S., Aryal, K., Gupta, M., & Bertino, E. (2026). *A Survey of Agentic AI and Cybersecurity: Challenges, Opportunities and Use-case Prototypes*.

Lupinacci, M., Pironti, F. A., Blefari, F., Romeo, F., Arena, L., & Furfaro, A. (2025). *The Dark Side of LLMs: Agent-based Attacks for Complete Computer Takeover* (arXiv:2507.06850). arXiv

Maynard, A. (2026). *The AI Cognitive Trojan Horse: How Large Language Models May Bypass Human Epistemic Vigilance*. https://doi.org/10.48550/arXiv.2601.07085

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.*, *1*(4), 26:1-26:45.

Panchal, N. (2025). What Are the Must-Know Agentic Design Patterns for 2026? *ProCreator - A Global UI UX Design Agency*.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5), 100988.

Ransbotham, S., Kiron, D., Khodabandeh, S., Iyer, S., & Das, A. (2025). The Emerging Agentic Enterprise: How Leaders Must Navigate a New Age of AI. *MIT Sloan Management Review*.

Räz, T., Pahud De Mortanges, A., & Reyes, M. (2025). Explainable AI in medicine: Challenges of integrating XAI into the future clinical routine. *Frontiers in Radiol.*, *5*

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & SOCIETY*, *41*(1), 259–278.

Rosenbacke, R., Rosenbacke, C., Rosenbacke, V., & McKee, M. (2025). *Beyond Hallucinations: The Illusion of Understanding in Large Language Models* (arXiv:2510.14665). arXiv.

Sahebi, S., & Formosa, P. (2025). The AI-mediated communication dilemma: Epistemic trust, social media, and the challenge of generative artificial intelligence. *Synthese*, *205*(3), 128.

Salehi, S., Singh, Y., Horst, K. K., Hathaway, Q. A., & Erickson, B. J. (2025). Agentic AI and Large Language Models in Radiology: Opportunities and Hallucination Challenges. *Bioengineering (Basel)*, *12*(12), 1303. (190544253).

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and challenges. *Information Fusion*, *126*, 103599.

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. *Scientific Reports*, *14*(1), 31150.

Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Petridis, S., Peng, Y.-H., Qiwei, L., Si, C., Xie, Y., Bigham, J. P., Bentley, F., Chai, J., Lipton, Z. C., Mei, Q., Terry, M., Yang, D., Morris, M. R., … Jurgens, D. (2025, October 29). *Position: Towards Bidirectional Human-AI Alignment*. The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track.

Shin, Y. (2025). Toward Human-Centered Artificial Intelligence for Users' Digital Well-Being: Systematic Review, Synthesis, and Future Directions. *JMIR Human Factors*, *12*, e69533.

Sidra, S., & Mason, C. (2025). Generative AI in Human-AI Collaboration: Validation of the Collaborative AI Literacy and Collaborative AI Metacognition Scales for Effective Use. *International Journal of Human–Computer Interaction*, 1–25.

Strebinger, A., & Treiblmaier, H. (2025). Eastern Ease and Western Worries? How collectivism and belief in magic increase user acceptance of operationally autonomous blackbox technologies. *Int. Journal of Information Management*, *84*, 102939.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24.

Tikhomirov, L., Semmler, C., McCradden, M., Searston, R., Ghassemi, M., & Oakden-Rayner, L. (2024). Medical artificial intelligence for clinicians: The lost cognitive perspective. *The Lancet Digital Health*, *6*(8), e589–e594.

Verma, S., Agarwal, A., Mruthyanjaya, P., Maharana, U., Mandal, M., Padhan, P., & Ahmed, S. (2025). Right Diagnoses with the Wrong Justification: Limitations of Current Large-Language Models for Screening of Rheumatoid Arthritis. *Journal of Clinical Rheumatology and Immunology*, *25*(1), 180–181. (72379971).

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(4), 589–606.

WEF (2025, October 15). *Non-human identities: Agentic AI's new frontier of cybersecurity risk*. World Economic Forum.

Yu, Y., Gomez-Cabello, C. A., Haider, S. A., Genovese, A., Prabha, S., Trabilsy, M., Collaco, B. G., Wood, N. G., Bagaria, S., Tao, C., & Forte, A. J. (2025). Enhancing Clinician Trust in AI Diagnostics: A Dynamic Framework for Confidence Calibration and Transparency. *Diagnostics*, *15*(17).

Zhang, X., Chen, Y., Yeh, S., & Li, S. (2025, October 29). *MetaMind: Modeling Human Social Thoughts with Metacognitive Multi-Agent Systems*. The Thirty-ninth Annual Conference on Neural Information Processing Systems.