Workshop paper presented at the Workshop on Human Agency and Skill in AI-Supported Work, held in conjunction with the 5th Annual Symposium on Human-Computer Interaction for Work (CHIWORK'26) in Linz, Austria, on June 22, 2026.

Open Questions Towards Skill-Sustaining Reliance in Reflective AI Engagement

Sander de Jong
Aalborg University
Aalborg, Denmark
sanderdj@create.aau.dk

Abstract

As AI systems are increasingly integrated into professional work, reflection strategies such as cognitive forcing and prompts that foster critical engagement have shown promise in reducing overreliance and improving decision quality. However, these strategies have primarily been evaluated as short-term interventions within single sessions. The next challenge is to assess whether such mechanisms sustain human agency and expertise over time. Drawing on prior work in AI-assisted decision-making, metacognition, and reflective AI engagement, we examine the challenges of designing and evaluating reflective mechanisms for long-term skill sustainability, considering individual differences in how users engage with such support, the organisational conditions under which it is implemented, and the gap between short-term evidence and long-term claims. We introduce open questions for the research community about the conditions under which reflective AI engagement can be sustained in practice.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Human-AI Interaction, Decision-Making, Large Language Models, Reflection

1 Introduction

Research on AI-supported professional work has extensively studied AI-assisted decision-making in single-sessions, focusing on how AI suggestions and explanations can improve human-AI decision-making [19, 20, 31, 32]. However, these strategies have primarily been evaluated as short-term interventions for improving task performance, trust calibration, or appropriate reliance. Moreover, in these sessions, many studies did not achieve complementary AI assistance, in which the human-AI team outperforms each acting individually [3, 29]. One of the underlying reasons is that people do not rely appropriately on AI assistance [21]. Strategies such as pre-commitment, uncertainty communication, and added reflection steps through cognitive forcing have been shown to mitigate overreliance and improve decision quality [5, 7, 9, 14, 24, 30]. As these strategies tend to require additional effort from workers, the challenge is to sustain this level of reflective engagement in real-world workflows.

More recently, LLM-generated rationales have been explored to help users understand and reflect on AI decisions [10, 18, 22]. However, their fluency can make incorrect outputs appear credible [4, 26], which may discourage the critical engagement they were meant to support. In group settings, the persuasiveness of LLMs may amplify majority viewpoints and increase conformity effects [8]. In contrast, when effectively used, LLMs can durably reduce belief in conspiracy theories [6] or help people understand and find common ground across opposing views [2, 28]. These dynamics suggest that the influence of LLM-generated rationales extends beyond individual decision-making. In group interactions, AI-mediated persuasion may shape shared professional norms within organisations.

Whether at the level of individual reasoning or collective deliberation, these findings raise a deeper concern. If the mechanisms designed to support human oversight can also undermine it, their value cannot be assessed solely by immediate decision quality. The question is whether professionals who rely on reflection mechanisms through human-AI interaction still develop their own expertise or slowly lose the capacity for independent judgement. For example, clinicians who routinely review AI-suggested diagnoses may perform well in these sessions, but are they still developing the diagnostic reasoning that unassisted practice would have demanded? A recent literature review of AI-induced deskilling in healthcare raises similar concerns across clinical specialities [23], arguing that strategies to monitor and mitigate skill erosion need to be developed at both the individual workflow and organisational levels.

In this paper, we discuss four challenges for sustaining human agency and expertise through reflective human-AI interaction, and raise open questions for the research community about the conditions under which such approaches can succeed in practice.

2 Varying Effects of Reflective Support

Reflective interaction mechanisms assume that prompting users to reason independently will improve their engagement with AI output. Empirical evidence suggests this varies across users. Bucinca et al. found that cognitive forcing reduced overreliance on average, but the benefit was most prominent among participants with higher Need for Cognition (NFC) [5]. De Jong et al. found similar effects of NFC on cognitive forcing and additionally showed that engagement varied across task difficulties, with participants engaging with cognitive forcing when they felt capable of reasoning independently, but relying on the AI suggestions otherwise [9]. This is consistent with Vasconcelos et al., who found that users weigh the effort of processing an explanation against its perceived benefit [30]. Overall, these findings suggest that effectiveness does not only depend on the design of the interaction but also on the user's expertise, domain familiarity, and personal characteristics.

In practice, these differences may complicate the design of decision support and raise questions about professional autonomy. If a reflective mechanism is perceived as ineffective or unnecessary,

professionals may simply ignore it. A novice professional who is still building foundational knowledge may benefit from a mechanism that withholds AI suggestions, as it allows them to develop independent reasoning. However, an expert may be able to directly evaluate AI advice critically, making withholding feel tedious instead of productive. One approach is to make reflective scaffolding adaptive. However, adaptation requires the system to assess the user's competence, which reduces the professional's autonomy over their own development. Over-scaffolding may unnecessarily constrain experienced professionals, while under-scaffolding may fail to support those still developing independent judgement.

Q1: When does reflective scaffolding undermine autonomy? Mechanisms that prompt independent reasoning assume the user requires support to think critically. For experienced professionals, this assumption may be unfounded and potentially undermining. How can reflective mechanisms support skill sustainability without implying that the user lacks competence? Who decides which users receive reflective scaffolding: the system, the organisation, or the professionals themselves?

3 Reflection in Efficiency-Driven Workflows

A primary driver of AI adoption in professional work is to enable faster decision-making and higher throughput while reducing cognitive load. Reflective interaction mechanisms, by design, contradict this promise. They introduce friction, ask users to slow down, and demand cognitive effort at moments where the AI could have provided an immediate answer [5, 9, 14, 24]. A controlled study can accommodate this trade-off. However, it is more problematic in a real-world environment that measures performance by quantity or speed.

If organisations do not recognise reflective engagement as an investment in long-term competence, these mechanisms are unlikely to be sustained in practice. Professionals using AI decision support typically need to reach productivity targets that leave little room for deliberate reflection. When reflective interactions add time to each decision and a professional makes numerous decisions per day, the cost becomes difficult to justify. This tension is particularly evident in high-stakes, risk-averse domains such as healthcare, where reflection and skill retention are closely linked to patient safety. This challenge extends beyond individual workplaces. When competing tools offer the same AI support without the added friction, organisations face pressure to adopt the faster alternative, making skill-sustaining designs harder to justify in the market.

Q2: How can skill-sustaining reliance withstand organisational and competitive pressure? Reflective interaction mechanisms take time. In workplaces that optimise for throughput, tools that deliberately slow professionals down may be difficult to justify, even if they preserve long-term competence. This challenge extends to the competitive landscape. If rival tools offer equivalent AI support that prioritises speed over reflection, skill-sustaining designs become a competitive disadvantage. Is this a challenge for interaction design, organisational policy, governmental policy, or a combination of these?

4 Sustaining Independent Metacognition

Reflective interaction mechanisms are designed to keep users cognitively engaged with AI output. For example, Fischer et al. describe a taxonomy of Socratic questions to promote critical reflection in AI-assisted clinical decision-making [12]. The need for such prompts is supported by Fernandes et al., who found that LLMs improved task performance but made users overconfident about their own accuracy [11]. Most participants relied on single prompts without further follow-up, suggesting that the LLM interaction did not encourage reflection. Rather than supporting metacognition, the AI replaced the opportunity to develop it naturally. Steyvers and Peters argue more broadly that LLMs impose substantial metacognitive demands on users but are reluctant to express uncertainty [27]. Because humans rely heavily on linguistic cues of uncertainty, the absence of doubt can lead to unwarranted reliance, further weakening the user's capacity for self-monitoring.

Reflective scaffolding, such as structured questions or prompted reflection, may address this gap in the short term. However, when the same prompts are encountered repeatedly, their effect may wear off through habituation as users learn to respond to the prompt without critically engaging with it. This is consistent with broader findings on warning fatigue, where repeated alerts are automatically dismissed [1]. This may cause scaffolding itself to become the source of engagement rather than the user's own metacognitive habits. A professional who only reflects on AI output when prompted by the system is not necessarily developing the capacity to reflect without that prompt. Over time, the scaffolding may replace the metacognitive capacity it was meant to protect. Removing or changing the tool leaves the professional without both AI support and the self-monitoring habits that independent practice would have developed.

Q3: Can skill-sustaining reliance itself become a form of deskilling? If a professional relies on an AI tool's reflective prompts to maintain their reasoning practice, the mechanism may replace the professional's own metacognitive habits rather than strengthen them.

5 From Short-Term Evidence to Long-Term Claims

Q3 addresses a limitation that extends beyond reflective scaffolding. Most empirical work on AI-assisted decision-making evaluates interventions within single sessions, making it impossible to determine whether their effects persist. Metrics such as task accuracy, reliance rate, and trust calibration are measured immediately after exposure. Therefore, the conclusions about reflection mentioned in the introduction only address session-level consequences.

A growing body of work has begun to examine how trust and reliance develop over repeated interactions. Kahr et al. studied trust development across repeated legal decision-making tasks, showing that trust is dynamic, shaped by AI accuracy, explanation type, and the timing of errors [16, 17]. Riedl and Bogert tracked chess players' use of AI feedback over time, finding that how people chose to engage with AI feedback determined whether they improved or stagnated [25]. In a field study with logistics professionals, Kahr et al. found that experts developed trust over time despite recognising AI imperfections, suggesting that apart from performance,

long-term trust also depends on transparency, system consistency, and professionals' interpersonal values [15]. These studies are an important step towards longitudinal evaluation. However, they focus on whether users appropriately rely on the AI over time, not on whether their capacity for independent reasoning is maintained or reduced through repeated interaction.

This distinction matters because the most relevant outcomes for skill sustainability may not be visible in trust or reliance measures. As discussed earlier, AI support can improve task performance while reducing users' ability to judge their own accuracy [11]. Similarly, Goh et al. found that giving physicians access to an LLM did not improve their diagnostic reasoning, even though the LLM outperformed them [13], suggesting that the interaction did not translate AI capability into better human reasoning. If professionals can use a highly capable AI without improving their own judgement, it raises the question of what happens to that judgement over months of repeated use.

Q4: How to navigate the asymmetry between short-term evidence and long-term claims? Reflective mechanisms have been shown to activate reasoning in short-term controlled sessions. However, it is unclear whether they sustain expertise longitudinally. How should the field act on promising short-term findings when the long-term evidence is insufficient? What level of evidence is sufficient to justify deploying reflective mechanisms in professional practice, and what level is needed to justify deciding against their deployment?

6 Conclusion

The four questions raised in this paper share a common underlying challenge. Reflective AI engagement is a promising direction, but its long-term viability depends on conditions that extend beyond interaction design. Whether reflective mechanisms sustain expertise depends on who uses them, whether organisations support them, and whether they build lasting capacity. We hope these questions serve as a starting point for discussion on the sustainability of human agency and expertise in AI-supported work environments.

References

[1] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2883–2892. doi:10.1145/2702123.2702322

[2] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale. *Proceedings of the National Academy of Sciences* 120, 41 (Oct. 2023), e2311627120. doi:10.1073/pnas.2311627120

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3411764.3445717

[4] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 152–163. doi:10.1609/icwsm.v18i1.31304

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 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 (April 2021), 188:1–188:21. doi:10.1145/3449287

[6] Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably Reducing Conspiracy Beliefs through Dialogues with AI. *Science* 385, 6714 (Sept. 2024), eadq1814. doi:10.1126/science.adq1814

[7] Sander de Jong, Maarten W. Bos, Niels van Berkel, and Maarten H. Lamers. 2025. Algorithm Appreciation or Aversion: The Effects of Accuracy Disclosure on Users' Reliance on Algorithmic Suggestions. *Behaviour & Information Technology* 0, 0 (Aug. 2025), 1–20. doi:10.1080/0144929X.2025.2535732

[8] Sander de Jong, Rune Møberg Jacobsen, Joel Wester, Senuri Wijenayake, Jorge Goncalves, and Niels van Berkel. 2025. Impact of Agent-Generated Rationales on Online Social Conformity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 3370–3384. doi:10.1145/3715275.3732217

[9] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW048:1–CSCW048:30. doi:10.1145/3710946

[10] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. arXiv:2402.17385 [cs] doi:10.48550/arXiv.2402.17385

[11] Daniela Fernandes, Steeven Villa, Salla Nicholls, Otso Haavisto, Daniel Buschek, Albrecht Schmidt, Thomas Kosch, Chenxinran Shen, and Robin Welsch. 2026. AI Makes You Smarter but None the Wiser: The Disconnect between Performance and Metacognition. *Computers in Human Behavior* 175 (Feb. 2026), 108779. doi:10.1016/j.chb.2025.108779

[12] Simon W. S. Fischer, Hanna Schraffenberger, Serge Thill, and Pim Haselager. 2025. A Taxonomy of Questions for Critical Reflection in Machine-Assisted Decision-Making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 1 (Oct. 2025), 940–954. doi:10.1609/aies.v8i1.36602

[13] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Josephine Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. 2024. Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. *medRxiv: The Preprint Server for Health Sciences* (March 2024), 2024.03.12.24303785. doi:10.1101/2024.03.12.24303785

[14] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 50:1–50:24. doi:10.1145/3359152

[15] Patricia Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2025. Good Performance Isn't Enough to Trust AI: Lessons from Logistics Experts on Their Long-Term Collaboration with an AI Planning System. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3706598.3713099

[16] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C.P. Snijders. 2023. It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 528–539. doi:10.1145/3581641.3584058

[17] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* 14, 4 (Dec. 2024), 29:1–29:30. doi:10.1145/3686164

[18] Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3714020

[19] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. doi:10.1145/3593013.3594087

[20] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

[21] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[22] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025. From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3706598.3713133

[23] Chiara Natali, Luca Marconi, Leslye Denisse Dias Duran, Massimo Miglioretti, and Federico Cabitza. 2025. AI-Induced Deskilling in Medicine: A Mixed Method

Literature Review for Setting a New Research Agenda. social science research network:5166364 doi:10.2139/ssrn.5166364

[24] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 102:1–102:15. doi:10.1145/3359204

[25] Christoph Riedl and Eric Bogert. 2026. Who Benefits from AI? Self-Selection, Skill Gap, and the Hidden Costs of AI Feedback. *Papers* 2409.18660 (April 2026).

[26] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the Conversational Persuasiveness of GPT-4. *Nature Human Behaviour* (May 2025), 1–9. doi:10.1038/s41562-025-02194-6

[27] Mark Steyvers and Megan A. K. Peters. 2025. Metacognition and Uncertainty Communication in Humans and Large Language Models. *Current Directions in Psychological Science* (Nov. 2025), 09637214251391158. doi:10.1177/09637214251391158

[28] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science* 386, 6719 (Oct. 2024), eadq2852. doi:10.1126/science.adq2852

[29] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour* 8, 12 (Dec. 2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[30] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 129:1–129:38. doi:10.1145/3579605

[31] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. doi:10.1145/3397481.3450650

[32] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852