How to Measure Cognitive Engagement in Machine-Assisted Decision-Making?

Simon W.S. Fischer[1,*], Hanna Schraffenberger[2]

[1]Donders Institute for Brain, Cognition, and Behaviour, Dpt. Human-Centred Intelligent Systems, Nijmegen, The Netherlands
[2]iHub, Radboud University, Nijmegen, The Netherlands

Abstract

Decision-support systems are used in various sectors, yet harbour the risks of overreliance and deskilling. To mitigate these risks, various interaction methods have been proposed that encourage the *cognitive engagement* of the decision-maker. However, there is currently no simple method to assess cognitive engagement during machine-assisted decision-making. We therefore propose the development of a self-report scale for cognitive engagement and offer a starting point for this. We present an overview of existing related scales from the education and healthcare sectors, and based on those, suggest possible items for such a new cognitive engagement scale. While future work is needed to finalise and validate the scale, it could ultimately serve as an evaluation instrument and guide the design and development of future decision-support systems that promote cognitive engagement in the form of critical thinking and reflection rather than overreliance and deskilling.

Keywords

Cognitive engagement, Reflection, Critical thinking, Evaluation, Overreliance, Deskilling, Decision-support system

1. Introduction

Decision-support systems (DSS) are widely used in various sectors, such as healthcare, education, and law, to assist decision-makers in making decisions. Studies show, however, that operators tend to accept incorrect recommendations [1, 2, 3], a phenomenon known as overreliance [4]. One consequence of overreliance on DSS or similar aids can be a reduction in critical thinking skills [5, 6, 7], often referred to as deskilling [8].

To mitigate overreliance and support effective human oversight, as required by the European AI Act, different approaches to human-AI interaction are proposed in the literature [9, 10, 11]. Specifically, frictional design strategies [12] aim to promote cognitive engagement of the decision-maker in the form of critical thinking and reflection during the decision-making process [13, 14, 15, 16, 17]. One prototype, for example, presents evidence for and against a decision, rather than providing a recommendation that needs to be accepted or rejected [18]. Although such approaches seem promising, evaluating their effectiveness in promoting deliberate decision-making through cognitive engagement remains a challenge.

The challenge is due to a lack of methods to assess cognitive engagement in machine-assisted decision-making, especially the entire decision-making process. Current evaluation methods of DSSs primarily focus on decision accuracy and the effect of machine recommendations on the final decision [19]. Available methods for measuring cognitive engagement, on the other hand, either (1) present technical barriers and require considerable effort and expertise, such as electroencephalography (EEG) [20, 21] or pupillometry [22], or (2) are not directly applicable to the context of machine-assisted decision-making, such as accessible self-report scales from the education domain [23, 24, 25].

To address this limitation, we propose a starting point for the development of a self-report scale that measures cognitive engagement in machine-assisted decision-making. We envision that our preliminary

HHAI-WS 2025: Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 9–13, 2025, Pisa, Italy
*Corresponding author.
simon.fischer@donders.ru.nl (S. W.S. Fischer)
0000-0003-2992-6563 (S. W.S. Fischer); 0000-0003-1847-2754 (H. Schraffenberger)

work will turn into a robust scale that is system-independent and thus widely applicable. The benefits of a self-report scale are its ease of use and low costs. Furthermore, factors associated with cognitive engagement, such as reflection, self-efficacy, and perceived control, are highly subjective, which makes a self-report scale a suitable measurement method. In view of this, "self-report scales are the most common approach to assessing cognitive engagement" in educational studies [24, p.66].

We see three main applications for such a cognitive engagement scale. First, it can be utilized to evaluate decision-support systems or similar aids and interventions. In this way, the scale could be a (proxy) instrument for measuring the extent of overreliane on DSS and thus also effective human oversight. Second, it can help design (future) interventions and interfaces by taking into account the scale items and aiming to achieve high scores on those. Third, the scale could serve as an intervention and checklist that is consulted by the decision-maker, e.g., to ensure that they also consider alternatives. In line with this, the scale could also be used to increase AI literacy, i.e., to sensitize operators to potential overreliance, as well as inform them about the importance of reflecting on and evaluating provided information.

According to Boateng et al. [26], there are three phases in the development of a scale. First, identifying and generating items. Second, constructing the scale, including pre-testing questions, and reducing the number of items. Third, assessing the reliability and validity of the scale. In this short paper, we focus on the initial phase of identifying and generating relevant items. To this end, we provide a collection of relevant existing scales (Table 1), particularly from the domains of education and healthcare, that are used to measure factors associated with cognitive engagement, such as reflection and decision self-efficacy. We suggest that these scales provide valuable input and that some of their items can be adapted and reused for our purposes (Table 2).

2. Background

We will briefly discuss the current focus of evaluation methods (2.1), argue for the need of a cognitive engagement scale (2.2), mention factors related to cognitive engagement (2.3), and point out approaches that aim to promote cognitive engagement (2.4).

2.1. Decision Outcome as the Current Focus of Evaluations

A textual analysis of 100 highly cited machine learning papers identified 59 values in the research field and concludes that [27, p.176]:

> The dominant values that emerged from the annotated corpus are: *Performance*, Generalization, Building on past work, Quantitative evidence, *Efficiency*, and Novelty. *(emphasis added)*

By focusing on model performance, accuracy, and efficiency, both during model development and model evaluation, only the outcome, i.e., the prediction, recommendation, or decision, is considered.

Similarly, current evaluations of reliance on DSSs focus on the decision outcome. In order to assess the effectiveness of an intervention, an initial decision without aid is usually compared with a decision made with the intervention [28, 29, 19]. Reverberi et al. [30], for example, offer formulas for calculating the influence of a DSS on the operator's decision, or the effect of a DSS on diagnostic accuracy. This formal and statistical approach and the focus on the outcome, however, are not sufficient to assess cognitive engagement during decision-making.

The first steps towards shifting this focus are presented in a study on interventions for critical thinking in AI-assisted knowledge work [31]. The authors developed "a *de novo* questionnaire to capture reflective thinking behaviours with AI-assisted workflows, modelled after Kember et al. [32]" [31, p.15]. This AI Workflow Reflective Thinking Questionnaire contains 14 items, including "I checked the AI suggestions for errors", "I considered the possibility that the AI suggestion could be wrong", or "I was critical or sceptical of the AI suggestions". While this questionnaire is commendable, it still places too much emphasis on the recommendation made, i.e., the outcome or result of the DSS.

In view of the above, we contend that a stronger focus on the decision-making *process* is required.

2.2. Shifting the Focus to the Decision-Making Process

Miller [18] proposes six criteria for good decision support based on the cardinal decision issues [33]. Accordingly, a decision-support system should help identify *options*, possible outcomes, i.e., *possibilities*, *values*, as well as help *judge* outcomes, weigh *trade-offs*, and be *understandable*. These criteria could serve as evaluation criteria. For example, the extent to which the decision support-system helps to identify stakeholder values could be assessed.

In line with this, a "good" decision, as we understand it, consists of the decision-maker being able to provide their own (justified) reasons for a course of action [34]. To this end, the decision-maker must analyse and evaluate information (provided by a DSS), weigh options, scrutinise assumptions and assess consequences of a decision [35]. The associated cognitive processes for these tasks include retrieving, understanding, analysing, and evaluating information, all of which require cognitive engagement [25]. It is therefore important to consider the entire decision-making process and the cognitive engagement in this process, and not just the result of the decision [36].

2.3. Factors Related to Cognitive Engagement

A predictor for cognitive engagement is *self-efficacy* [23], which is the ability to produce desired outcomes by one's own actions [37]. Relying on own skills can strengthen *confidence* in one's own decision-making. Higher confidence is associated with better clinical decision-making [38], more critical thinking, and less reliance on decision-support systems [7]. Moreover, confidence has an effect on *motivation*, which, in turn, is a modulator and precursor to cognitive engagement [23, 39].

Closely linked to self-efficacy is *sense of agency* [40] and "beliefs in personal control" [41]. Accordingly, for the decision-maker to have a sense of agency or (perceived) control over the decision-making process and thus rely less on a DSS, the decision-support system should ideally support them in their own reasoning so that they are confident and have fewer doubts about their judgments.

Metacognitive processes related to (decision) self-efficacy and cognitive engagement are *deliberation* and *reflection* [42, 43]. Deliberation is the act of weighing options and making decisions carefully, while reflection examines held assumptions and tacit knowledge [44]. Reflection has been shown to enhance critical thinking and reasoning, as well as improve decision-making [45, 46, 47, 48, 17, 49].

2.4. Approaches to Promote Cognitive Engagement

Compared to the prevailing approach of providing recommendations or predictions based on input data, several prototypes aim to support the decision-maker by promoting the previously mentioned factors, such as self-efficacy, sense of agency, or reflection.

As mentioned earlier, a so-called evaluative AI presents information for and against a decision, allowing the decision-maker to make their own decision [18]. Similarly, a so-called Judicial DSS provides explanations for opposing decision options [9]. In this case, the DSS provides two explanations, one for the presence and one for the absence of spinal fractures on an X-ray image. By doing so and not making a recommendation, the Judicial AI is "preserving a sense of agency" [9].

In line with the provision of opposing information, it is argued that "computational tools need to be intentionally designed in such a way that they actively support critical reflection" [50, p.3]. One example is a prototype from the financial domain, based on a large language model [51]. The chatbot asks investors for their investment rationale and provides feedback in the form of indications of blind spots and additional considerations to help them reflect on their reasoning. Another study presents knowledge workers conducting AI-assisted shortlisting tasks with provocations, i.e., brief textual prompts that offer critiques and alternatives to machine recommendations [31]. The aim of these provocations is to induce critical thinking. Besides, the authors also developed the above-mentioned *AI Workflow Reflective Thinking Questionnaire*, which we will return to shortly.

Although increasing, the number of these approaches to promoting cognitive engagement is still relatively small compared to conventional DSSs. We believe that by shifting the focus of evaluation from decision outcome and accuracy to the decision process, relevant decision-support systems can be designed that go beyond the mere presentation of a final recommendation (section 5).

3. Methods

For our envisioned cognitive engagement scale, we drew some general inspiration from the Technology Acceptance Model (TAM) [52], and the Usability and Ease of Use Questionnaire (USE) [53]. These self-report scales are widely used in human-computer interaction and measure the acceptance and usability of a technology through items, such as "It helps me to be more effective". Similarly, a possible item for the cognitive engagement scale could be: "It helps me to reflect" (see Table 2), where the level of agreement can range from "extremely disagree" to "extremely agree".

While TAM and USE provided some initial ideas for the content of our scale, we searched for literature more closely related to our focus, specifically scales measuring cognitive engagement in machine-assisted decision-making. The only relevant scale we found was the aforementioned AI Workflow Reflective Thinking Questionnaire [31]. The non-validated questionnaire goes in a similar direction to what we envision, so we adopted items from it (Table 2). We nevertheless found that it still focuses too much on the decision outcome and only partially assesses cognitive engagement.

From the educational domain, two reviews on measuring cognitive engagement proved to be useful [24, 23]. Although most of the listed scales are too context-specific and thus not directly applicable to our use case, they served as a starting point. In addition, these reviews discuss the relationship between cognitive engagement and various (psychological) factors as well as metacognitive tasks such as motivation, self-efficacy, self-regulation, and self-reflection, as we mentioned before (section 2.3).

To identify other potentially relevant items and scales, we used the factors associated with cognitive engagement to search for individual scales. The generic search query was *"keyword + (scale OR questionnaire OR measure)"*, where keywords were cognitive engagement, critical thinking, reflection, deliberation, self-efficacy, decision-autonomy, motivation, and confidence.

As an inclusion criterion, we used the perceived usability and applicability of the scales to assess the extent of the decision-maker's cognitive engagement during machine-assisted decision-making and the extent to which the DSS contributed to promoting cognitive engagement. Accordingly, the scales had to be somewhat generic. Some scales were too specific to other domains or tasks and thus less useful. For example, a scale to measure the level of confidence in nursing students contains items like "Dealing with upset of angry relatives" [54], or a scale to measure reflection competencies of clinical nurses asks "I think about what I can do for patients in addition to my assigned tasks" [55]. Moreover, we decided to only include publicly available scales.

We reused and adapted items from the collected scales to create relevant items (Table 2). The adaptation was quite simple. For example, we replaced "facts about medication choices" with "recommendation". Furthermore, we replaced context-specific information, e.g., "ideas in course material" with more general "information" or "DSS recommendation". We have also added context where necessary to make the items more suitable for the task of decision-making.

4. Collection of Self-Report Scales

We provide a collection of relevant existing scales that can help identify and generate items for a scale that assesses cognitive engagement in machine-assisted decision-making (Table 1). We will discuss scales for measuring cognitive engagement in education [25], as well as various scales on factors related to cognitive engagement, such as decision self-efficacy, motivation, deliberation, and reflection.

In addition, we provide some example items to show how existing scales can function as inspiration and how items can be reused or adapted to create potential items for a cognitive engagement scale in machine-assisted decision-making (Table 2). From these example items, potential dimensions of a scale

Table 1
A non-exhaustive collection of scales for measuring cognitive engagement and associated factors.

Scale	Description
AI Workflow Reflective Thinking Questionnaire [31]	14-item 5-point Likert scale to capture reflective thinking behaviours in AI-assisted workflows.
Cognitive Engagement with Technology [25]	10-item 5-point scale to assess how students use technology to perform cognitive learning activities (retrieving, processing, generating), building on Bloom's Digital Taxonomy.
Cognitive Strategy Use and Motivation [23]	41-item scale to assess academic achievement. Items are categorised into type and degree of cognitive strategy use (deep/shallow, with motivation as modulator), use of self-regulatory processes (goal setting, planning, monitoring learning, and self-reflection and reaction), and degree of effort exerted.
Decision Self-Efficacy [56]	11-item 5-point scale to measure self-confidence or beliefs in one's abilities in decision-making, including shared decision-making.
DelibeRATE [57]	9-item 7-point scale assessing to which extent patients were willing to make a decision about the surgery to be chosen.
Groningen Reflection Ability Scale [58]	23-item on 5-point scale assessing personal reflection abilities (self-reflection, empathetic reflection, and reflective communication) of medical students.
Sense of Agency [41]	13-item scale to measure person's belief of having agency.

can be derived, such as understanding, confidence, evaluating information, comparing alternatives, and reflection.

The AI Workflow Reflective Thinking Questionnaire assesses the extent to which knowledge workers reflect on the recommendations provided by a decision-support system. The scale was specifically developed to evaluate a prototype proposed by the same authors, which, as mentioned above, provides provocations to recommendations [31]. Nevertheless, since a future scale should gauge the level of reflection, some items can be reused and adapted for our purposes.

The Cognitive Engagement with Technology (CET) Scale measures how students use technology for different cognitive tasks [25]. These cognitive tasks are derived from Bloom's Digital Taxonomy, which integrates the use of technology to the original Bloom's taxonomy [59]. In Bloom's taxonomy, cognitive processes are categorised into the following: remembering, understanding, applying, analysing, evaluating, and creating. In our context, a DSS should ideally support certain cognitive tasks. As will become apparent in Table 2, the dimensions understanding, analysing, and evaluating are likely to be of great importance in a future scale for cognitive engagement.

The Cognitive Strategy Use and Motivation Scale measures the motivation of students and its impact on cognitive engagement [23]. The scale consists of three dimensions, namely self-regulation, deep strategy use, and shallow strategy use. Especially the distinction between deep and shallow engagement could be helpful for a future scale. As the author notes, deep engagement is an active and intentional process, whereas shallow engagement involves "cognitive actions that are more mechanical than thoughtful (e.g., rote rehearsal and verbatim memorization strategies)" [23, p.15]. For our purposes, it seems desirable that decision-makers engage deeply with the information provided by a DSS and the other decision-relevant information, and that the DSS promotes this behaviour. Accordingly, a scale should capture the level of engagement.

The Decision Self-Efficacy Scale measures confidence in making an informed choice about available treatment options [56]. In our context, and as mentioned earlier, confidence is associated with less reliance on DSS and better decision-making. Moreover, higher confidence is likely to require deep engagement (previous scale) with the decision, the prevailing assumptions, and the possible consequences. A future scale for cognitive engagement should therefore measure how confident the decision-maker is.

The DelibeRATE Scale, developed for a medical context, assesses the deliberation process and "the extent to which participants were thinking about their decision" [57, p.212], where a higher score

Table 2
Examples of reusing and adapting existing scales and items to create relevant items to measure cognitive engagement in machine-assisted decision-making.

Scale	Sample Items	Adapted Items
AI Workflow Reflective Thinking Questionnaire [31]	• I updated my understanding of my own preferences. • I updated my understanding of the AI system and its limitations • I was critical or sceptical of the AI suggestions. • I considered the possibility that the AI suggestion could be wrong.	The DSS helped me to • be aware of my preferences/assumptions. • understand how the DSS works and what the limitations are. • evaluate the recommendation. • explore alternative solutions.
Cognitive Engagement with Technology [25]	Indicate how often you use technology to • compare multiple sources of information. • analyse different aspects of a problem or issue. • assess the quality of an online source.	The DSS helped me to • compare multiple sources/pieces of information. • analyse different aspects of a problem or decision. • assess the quality of a recommendation/prediction.
Cognitive Strategy Use and Motivation [23]	Deep Strategy Use Dimension • I compared and contrasted different concepts. • I evaluated the usefulness of the ideas presented in course materials. • I made sure I understood material that I studied.	The DSS helped me to • compare and contrast different options. • evaluate information. • understand how the recommendation was made.
Decision Self-Efficacy [56]	I feel confident that I can • get the facts about the medication choices available to me. • get the facts about the risks and side effects of each choice. • understand the information enough to be able to make a choice.	I am confident that I • understand each DSS recommendation. • understand the consequences and risks of the recommendation. • understand the recommendations enough to make a decision
DelibeRATE [57]	• I know enough about each option to help me decide. • I know about the advantages and disadvantages of each option. • I can judge which option is better for me.	• I know enough about each option to help me decide. • I know about the advantages and disadvantages of each option. • I can judge which option is better for me / the affected person (e.g., patient).
Groningen Reflection Ability Scale [58]	• I take a closer look at my own habits of thinking. • I want to know why I do what I do. • I am aware of the emotions that influence my thinking. • I can see an experience from different standpoints. • I reject different ways of thinking.	The DSS helped me to • assess my own habits of thinking. • understand why I am accepting/rejecting this recommendation. • become aware of the emotions / cognitive biases that influence my thinking. • to consider different viewpoints. • consider different ways of thinking/solutions.
Sense of Agency [41]	• I am in full control of what I do. • I am the author of my actions.	• I feel in control of the decision. • I feel I make the final decision (I am not merely following the recommendation). • The DSS supports my own reasoning.

indicates the readiness to decide which treatment option to choose. The more one thinks about their decision, the more likely they are to be confident about it. The DelibeRATE scale thus has some overlaps with the Decision Self-Efficacy Scale.

The Groningen Reflection Ability Scale (GRAS) measures the ability of medical students and doctors to reflect on their behaviour [58]. The authors distinguish between three cognitive-emotional levels of reflection, namely clinical reasoning, scientific reflection, and personal reflection. GRAS focuses on personal reflection, i.e., reflection on emotions, assumptions, and beliefs based on experiences, which, according to the authors, is a process of sense-making. Similarly, a future scale for cognitive engagement should take into account this sense-making process during decision-making and the role of cognitive biases in this process.

The Sense of Agency (SoA) Scale measures a person's belief that they have agency [41]. Although rather general, the scale can help to create items that inquire about the (perceived) control of the decision-maker. For example, items can ask about the "ownership" of a decision, which also implies that the decision-maker knows the reasons for an action and possible influencing factors instead of unthinkingly accepting a machine recommendation.

5. Discussion

We presented a starting point for the creation of relevant items for a scale to measure cognitive engagement in machine-assisted decision-making. To this end, we have compiled a collection of promising and relevant available scales for our context (Table 1). Since these scales are used in education or healthcare, we have also provided examples of how items could be reused and adapted (Table 2). As can be seen, various items overlap, e.g., from the Decision Self-Efficacy Scale and the DelibeRATE Scale. The next step in creating a scale thus requires the differentiation and reduction of items, and possibly the addition of further items not covered here. A challenge in this regard is to find a balance between a scale that is too generic and one that is too context-specific. It might be necessary to select different items from different scales, based on case-specific and context-dependent requirements.

In addition, the right balance (factor loading) and amount of items must be found. For example, confidence in a decision can be counterproductive to cognitive engagement if the decision-maker is overconfident and does not engage with the information or scrutinise their (biased) assumptions. In order to create a counterbalance to overconfidence, items are needed that, for example, ask about the degree of reflection and the consideration of alternatives [35]. It is important that items are designed in such a way that they take into account the entire decision-making process, and not just focus on the outcome of the decision or recommendation of the DSS. To delineate this focus, different dimensions of a scale could be helpful, such as understanding, confidence, evaluating information, comparing alternatives, and reflection, with the dimensions understanding and confidence emphasising the outcome, and the others the process.

The objective of the developed scale could be twofold, which leads to different formulations of the scale items. First, a scale could measure the extent to which a system or DSS promotes cognitive engagement in the form of critical thinking and reflection. Second, taking cognitive dispositions into account, it could measure the extent to which the decision-maker engages cognitively while interacting with a DSS, e.g., by evaluating information. As Greene [23, p.27] notes:

> [...] cognitive engagement is not a stable characteristic of either a learner or a learning environment but rather a fluid set of processes that can be influenced by learners themselves and by the environment.

In view of this, decision-support systems, i.e., the technological environment, can influence critical thinking and cognitive engagement as much as the decision-makers.

Scale items could therefore either be phrased as (1) "The DSS helped me to evaluate the recommendation" or (2) "I evaluated the machine recommendation". We assume the scale would need to address both dimensions. Nevertheless, some means or information are necessary to evaluate the machine

recommendation (2), which ideally is provided by the DSS. We therefore suggest that the focus of the scale should be on assessing the extent to which the decision *support*-system supports cognitive processes related to cognitive engagement (1). For this reason, we adapted the Groningen Reflection Ability Scale, which originally assesses personal reflection (2), and changed the items to assess the extent to which a DSS contributes to reflection (1).

With the development of a scale for cognitive engagement in machine-assisted decision-making, we also like to draw attention to the design of decision-support systems. As mentioned, DSSs typically emphasise the outcome of the decision by providing recommendations, while key evaluation metrics focus on decision accuracy, performance, and efficiency [27]. We believe that not only the evaluation but also the design of DSSs would benefit from more attention to the underlying decision-making process [36]. Ideally, future systems will support cognitive processes like deliberation and reflection, as well as promote psychological characteristics, such as motivation and self-efficacy.

In the meantime, a future scale itself could prove to be a useful intervention in the form of a reminder, helping decision-makers to become aware of the need to evaluate the DSS recommendation. In line with this, the scale could also be used for training purposes and support AI literacy.

6. Conclusion

Current evaluation methods of decision-support systems do not take into account cognitive engagement and are thus not sufficient to measure it. Yet, assessing cognitive engagement becomes increasingly important considering the overreliance on decision-support systems or similar aids and its consequence of deskilling. We have therefore provided a starting point for the first phase of scale development, namely the identification and creation of relevant items (Table 2). In the next steps, a complete scale would have to be constructed and then validated.

A future scale for cognitive engagement in machine-assisted decision-making has the potential to shift the focus from decision accuracy, performance, and efficiency, to more human reflection, deliberation, and critical thinking. Cognitive engagement can serve as a new evaluation metric and thus contribute to the design and development of (future) decision-support systems that support cognitive factors in the decision-making process instead of just providing final recommendations.

Acknowledgments

We thank Linus Holmberg, Pim Haselager, and three anonymous reviewers for their feedback on the extended abstract of this short paper. Thanks to the organisers and participants of the Frictional AI workshop for their questions and discussions on our presentation. Comments from Niels van Berkel have helped to (tentatively) clarify the difference between cognitive engagement and cognitive load, which will prove helpful for future work. This research is funded by the Donders Centre for Cognition.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] T. Dratsch, X. Chen, M. Rezazade Mehrizi, R. Kloeckner, A. Mähringer-Kunz, M. Püsken, B. Baeßler, S. Sauer, D. Maintz, D. Pinto dos Santos, Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance, Radiology 307 (2023) e222176. doi:10.1148/radiol.222176.

[2] M. Jacobs, J. He, M. F. Pradier, B. Lam, A. C. Ahn, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, K. Z. Gajos, Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A

Sociotechnical Lens, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2021, pp. 1–14. doi:10.1145/3411764.3445385.
[3] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, H. Kittler, Human–computer collaboration for skin cancer recognition, Nature Medicine 26 (2020) 1229–1234. doi:10.1038/s41591-020-0942-0.
[4] S. Passi, M. Vorvoreanu, Overreliance on AI: Literature Review, Technical Report MSR-TR-2022-12, Microsoft, 2022.
[5] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, R. Mariman, Generative AI Can Harm Learning, 2024. doi:10.2139/ssrn.4895486.
[6] M. Gerlich, AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking, Societies 15 (2025) 6. doi:10.3390/soc15010006.
[7] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, N. Wilson, The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2025, pp. 1–22. doi:10.1145/3706598.3713778.
[8] S. Vallor, Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character, Philosophy & Technology 28 (2015) 107–124. doi:10.1007/s13347-014-0156-9.
[9] F. Cabitza, L. Famiglini, C. Fregosi, S. Pe, E. Parimbelli, G. A. La Maida, E. Gallazzi, From Oracular to Judicial: Enhancing Clinical Decision Making through Contrasting Explanations and a Novel Interaction Protocol, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, ACM, Cagliari Italy, 2025, pp. 745–754. doi:10.1145/3708359.3712157.
[10] F. Cabitza, A. Campagner, L. Famiglini, C. Natali, V. Caccavella, E. Gallazzi, Let Me Think! Investigating the Effect of Explanations Feeding Doubts About the AI Advice, in: A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, volume 14065, Springer Nature Switzerland, Cham, 2023, pp. 155–169. doi:10.1007/978-3-031-40837-3_10.
[11] V. Danry, P. Pataranutaporn, Y. Mao, P. Maes, Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, ACM, Hamburg Germany, 2023, pp. 1–13. doi:10.1145/3544548.3580672.
[12] A. L. Cox, S. J. Gould, M. E. Cecchinato, I. Iacovides, I. Renfree, Design Frictions for Mindful Interactions: The Case for Microboundaries, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, San Jose California USA, 2016, pp. 1389–1397. doi:10.1145/2851581.2892410.
[13] C.-W. Chiang, Z. Lu, Z. Li, M. Yin, Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate, in: Proceedings of the 29th International Conference on Intelligent User Interfaces, ACM, Greenville SC USA, 2024, pp. 103–119. doi:10.1145/3640543.3645199.
[14] P. Haselager, H. Schraffenberger, S. Thill, S. Fischer, P. Lanillos, S. van de Groes, M. van Hooff, Reflection Machines: Supporting Effective Human Oversight Over Medical Decision Support Systems, Cambridge Quarterly of Healthcare Ethics 33 (2023) 380–389. doi:10.1017/S0963180122000718.
[15] S. Ma, Q. Chen, X. Wang, C. Zheng, Z. Peng, M. Yin, X. Ma, Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2025, pp. 1–23. doi:10.1145/3706598.3713423.
[16] A. Sarkar, AI Should Challenge, Not Obey, Communications of the ACM 67 (2024) 18–21. doi:10.1145/3649404.
[17] H. G. Schmidt, S. Mamede, Improving diagnostic decision support through deliberate reflection: A proposal, Diagnosis 10 (2023) 38–42. doi:10.1515/dx-2022-0062.
[18] T. Miller, Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support

using Evaluative AI, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, ACM, Chicago IL USA, 2023, pp. 333–342. doi:10.1145/3593013.3594001.

[19] F. Cabitza, A. Campagner, R. Angius, C. Natali, C. Reverberi, AI Shall Have No Dominion: On How to Measure Technology Dominance in AI-supported Human decision-making, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ACM, Hamburg Germany, 2023, pp. 1–20. doi:10.1145/3544548.3581095.

[20] M. Hassib, M. Khamis, S. Friedl, S. Schneegass, F. Alt, Brainatwork: Logging cognitive engagement and tasks in the workplace using electroencephalography, in: Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia, ACM, Stuttgart Germany, 2017, pp. 305–310. doi:10.1145/3152832.3152865.

[21] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, P. L. Craven, EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks, Aviation, Space, and Environmental Medicine 78 (2007).

[22] P. Van Der Wel, H. Van Steenbergen, Pupil dilation as an index of effort in cognitive control tasks: A review, Psychonomic Bulletin & Review 25 (2018) 2005–2015. doi:10.3758/s13423-018-1432-y.

[23] B. A. Greene, Measuring Cognitive Engagement With Self-Report Scales: Reflections From Over 20 Years of Research, Educational Psychologist 50 (2015) 14–30. doi:10.1080/00461520.2014.989230.

[24] S. Li, Measuring Cognitive Engagement: An Overview of Measurement Instruments and Techniques, International Journal of Psychology and Educational Studies 8 (2022) 63–76. doi:10.52380/ijpes.2021.8.3.239.

[25] V. W. Vongkulluksn, L. Lu, M. J. Nelson, K. Xie, Cognitive engagement with technology scale: A validation study, Educational technology research and development 70 (2022) 419–445. doi:10.1007/s11423-022-10098-9.

[26] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, S. L. Young, Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer, Frontiers in Public Health 6 (2018) 149. doi:10.3389/fpubh.2018.00149.

[27] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, M. Bao, The Values Encoded in Machine Learning Research, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, ACM, Seoul Republic of Korea, 2022, pp. 173–184. doi:10.1145/3531146.3533083.

[28] Z. Guo, Y. Wu, J. D. Hartline, J. Hullman, A Decision Theoretic Framework for Measuring AI Reliance, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM, Rio de Janeiro Brazil, 2024, pp. 221–236. doi:10.1145/3630106.3658901.

[29] A. Klingbeil, C. Grützner, P. Schreck, Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI, Computers in Human Behavior 160 (2024) 108352. doi:10.1016/j.chb.2024.108352.

[30] C. Reverberi, T. Rigon, A. Solari, C. Hassan, P. Cherubini, GI Genius CADx Study Group, G. Antonelli, H. Awadie, S. Bernhofer, S. Carballal, M. Dinis-Ribeiro, A. Fernández-Clotett, G. F. Esparrach, I. Gralnek, Y. Higasa, T. Hirabayashi, T. Hirai, M. Iwatate, M. Kawano, M. Mader, A. Maieron, S. Mattes, T. Nakai, I. Ordas, R. Ortigão, O. O. Zúñiga, M. Pellisé, C. Pinto, F. Riedl, A. Sánchez, E. Steiner, Y. Tanaka, A. Cherubini, Experimental evidence of effective human–AI collaboration in medical decision-making, Scientific Reports 12 (2022) 14952. doi:10.1038/s41598-022-18751-2.

[31] I. Drosos, A. Sarkar, Xiaotong, Xu, N. Toronto, "It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work, 2025. doi:10.48550/arXiv.2501.17247. arXiv:2501.17247.

[32] D. Kember, D. Y. P. Leung, A. Jones, A. Y. Loke, J. McKay, K. Sinclair, H. Tse, C. Webb, F. K. Yuet Wong, M. Wong, E. Yeung, Development of a Questionnaire to Measure the Level of Reflective Thinking, Assessment & Evaluation in Higher Education 25 (2000) 381–395. doi:10.1080/713611442.

[33] J. F. Yates, G. A. Potworowski, Evidence-Based *Decision* Management, in: D. M. Rousseau (Ed.), The Oxford Handbook of Evidence-Based Management, 1 ed., Oxford University Press, 2012, pp. 198–222. doi:10.1093/oxfordhb/9780199763986.013.0012.

[34] D. J. Coates, P. Swenson, Reasons-responsiveness and degrees of responsibility, Philosophical

Studies 165 (2013) 629–645. doi:10.1007/s11098-012-9969-5.
[35] S. W. S. Fischer, H. Schraffenberger, S. Thill, P. Haselager, A Taxonomy of Questions for Critical Reflection in Machine-Assisted Decision-Making, 2025. doi:10.48550/ARXIV.2504.12830.
[36] Z. T. Zhang, S. S. Feger, L. Dullenkopf, R. Liao, L. Süsslin, Y. Liu, A. Butz, Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–32. doi:10.1145/3687024.
[37] A. Bandura, Toward a Psychology of Human Agency, Perspectives on Psychological Science 1 (2006) 164–180. doi:10.1111/j.1745-6916.2006.00011.x.
[38] P. L. Hart, L. Spiva, N. Mareno, Psychometric Properties of the Clinical Decision-Making Self-Confidence Scale, Journal of Nursing Measurement 22 (2014) 312–322. doi:10.1891/1061-3749.22.2.312.
[39] M. Singh, P. James, H. Paul, K. Bolar, Impact of cognitive-behavioral motivation on student engagement, Heliyon 8 (2022) e09843. doi:10.1016/j.heliyon.2022.e09843.
[40] R. Legaspi, W. Xu, T. Konishi, S. Wada, N. Kobayashi, Y. Naruse, Y. Ishikawa, The sense of agency in human–AI interactions, Knowledge-Based Systems 286 (2024) 111298. doi:10.1016/j.knosys.2023.111298.
[41] A. Tapal, E. Oren, R. Dar, B. Eitam, The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment, Frontiers in Psychology 8 (2017) 1552. doi:10.3389/fpsyg.2017.01552.
[42] K. A. Lambe, G. O'Reilly, B. D. Kelly, S. Curristan, Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review, BMJ Quality & Safety 25 (2016) 808–820. doi:10.1136/bmjqs-2015-004417.
[43] S. Prakash, R. M. Sladek, L. Schuwirth, Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies, Medical Teacher 41 (2019) 517–524. doi:10.1080/0142159X.2018.1497786.
[44] D. A. Schön, The Reflective Practitioner: How Professionals Think in Action, Basic Books, New York, 1983.
[45] A. Ghanizadeh, The interplay between reflective thinking, critical thinking, self-monitoring, and academic achievement in higher education, Higher Education 74 (2017) 101–114. doi:10.1007/s10734-016-0031-y.
[46] B. J. Hess, R. S. Lipner, V. Thompson, E. S. Holmboe, M. L. Graber, Blink or Think: Can Further Reflection Improve Initial Diagnostic Impressions?, Academic Medicine 90 (2015) 112–118. doi:10.1097/ACM.0000000000000550.
[47] Z. Khoshgoftar, M. Barkhordari-Sharifabad, Medical students' reflective capacity and its role in their critical thinking disposition, BMC Medical Education 23 (2023) 198. doi:10.1186/s12909-023-04163-x.
[48] S. Mamede, H. G. Schmidt, Deliberate reflection and clinical reasoning: Founding ideas and empirical findings, Medical Education 57 (2023) 76–85. doi:10.1111/medu.14863.
[49] C. Walger, K. D. D. Roglio, G. Abib, HR managers' decision-making processes: A "reflective practice" analysis, Management Research Review 39 (2016) 655–671. doi:10.1108/MRR-11-2014-0250.
[50] K. Glinka, C. Müller-Birn, Critical-Reflective Human-AI Collaboration: Exploring Computational Tools for Art Historical Image Retrieval, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–33. doi:10.1145/3610054.
[51] L. Reicherts, Z. T. Zhang, E. Von Oswald, Y. Liu, Y. Rogers, M. Hassib, AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama Japan, 2025, pp. 1–19. doi:10.1145/3706598.3713295.
[52] F. D. Davis, Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, MIS Quarterly 13 (1989) 319. doi:10.2307/249008. arXiv:249008.
[53] A. M. Lund, Measuring Usability with the USE Questionnaire, Usability Interface 8 (2001) 3–6.
[54] P. Walsh, P. Owen, N. Mustafa, The creation of a confidence scale: The confidence in managing challenging situations scale, Journal of Research in Nursing 26 (2021) 483–496. doi:10.1177/

1744987120979272.
[55] S. Shin, E. Hong, J. Do, M. S. Lee, Y. Jung, I. Lee, Development of Critical Reflection Competency Scale for Clinical Nurses, International Journal of Environmental Research and Public Health 19 (2022) 3483. doi:10.3390/ijerph19063483.
[56] A. M. O'Connor, User Manual - Decision Self-Efficacy Scale, 1995. URL: http://decisionaid.ohri.ca/docs/develop/User_Manuals/UM_Decision_SelfEfficacy.pdf.
[57] S. Sivell, A. Edwards, A. S. Manstead, M. W. Reed, L. Caldon, K. Collins, A. Clements, G. Elwyn, Increasing readiness to decide and strengthening behavioral intentions: Evaluating the impact of a web-based patient decision aid for breast cancer treatment options (BresDex: Www.bresdex.com), Patient Education and Counseling 88 (2012) 209–217. doi:10.1016/j.pec.2012.03.012.
[58] L. C. Aukes, J. Geertsma, J. Cohen-Schotanus, R. P. Zwierstra, J. P. Slaets, The development of a scale to measure personal reflection in medical practice and education, Medical Teacher 29 (2007) 177–182. doi:10.1080/01421590701299272.
[59] L. Anderson, D. Krathwohl, A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Longman, New York, 2001.