

Supporting Reflection and Forward-Looking Reasoning With Data-Driven Questions

Simon W.S. Fischer

simon.fischer@donders.ru.nl

Donders Institute for Brain, Cognition, and Behaviour,
Radboud University
Dpt. of Human-Centred Intelligent Systems
Nijmegen, The Netherlands

Serge Thill

serge.thill@donders.ru.nl

Donders Institute for Brain, Cognition, and Behaviour,
Radboud University
Dpt. of Human-Centred Intelligent Systems
Nijmegen, The Netherlands

Hanna Schraffenberger

hanna.schraffenberger@ru.nl

Interdisciplinary Hub for Digitalization and Society (iHub)
and Institute for Computing and Information Sciences
(iCIS), Radboud University
Nijmegen, The Netherlands

Pim Haselager

pim.haselager@donders.ru.nl

Donders Institute for Brain, Cognition, and Behaviour,
Radboud University
Dpt. of Human-Centred Intelligent Systems
Nijmegen, The Netherlands

Abstract

Many generative AI systems as well as decision-support systems (DSSs) provide operators with predictions or recommendations. Various studies show, however, that people can mistakenly adopt the erroneous results presented by those systems. Hence, it is crucial to promote critical thinking and reflection during interaction. One approach we are focusing on involves encouraging reflection during machine-assisted decision-making by presenting decision-makers with data-driven questions. In this short paper, we provide a brief overview of our work in that regard, namely: 1) the development of a question taxonomy, 2) the development of a prototype in the medical domain and the feedback received from clinicians, 3) a method for generating questions using a large language model, and 4) a proposed scale for measuring cognitive engagement in human-AI decision-making. In doing so, we contribute to the discussion about the design, development, and evaluation of tools for thought, i.e., AI systems that provoke critical thinking and enable novel ways of sense-making.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods; HCI theory, concepts and models**; • **Information systems** → *Decision support systems*.

Keywords

questions, reflection, critical thinking, cognitive engagement, over-reliance, human-AI interaction

ACM Reference Format:

Simon W.S. Fischer, Hanna Schraffenberger, Serge Thill, and Pim Haselager. 2026. Supporting Reflection and Forward-Looking Reasoning With Data-Driven Questions. In *Proceedings of Tools For Thought (CHI2026 Workshop on Tools for Thought)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Generative AI as well as machine learning models are increasingly being used in decision-making by providing data-driven insights. Various studies show, however, that people tend to rely too much on the outputs or recommendations of these systems. A study by Anthropic, for example, identifies “disempowerment potential” of large language models where people tend to adopt the beliefs embedded in the language model [21]. Similarly, it has been shown that language models can influence moral judgement [12] and opinions [10]. In machine-assisted decision-making with so-called decision-support systems (DSSs), it has been found that even expert decision-makers, such as clinicians, are prone to accepting wrong recommendations from these systems [1, 3, 9]. This phenomenon is known as overreliance [17].

Particularly in high-risk sectors, such as healthcare, it is crucial that decision-makers, e.g., clinicians, are supported in such a way that they know when to accept a machine recommendation and when not to. Accordingly, legislation such as the European AI Act mandates *human oversight*, which requires systems to be designed and developed in such a way that persons are enabled “to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)” (Article 14, 4b). It is therefore important, and in some cases even necessary, to promote cognitive engagement and continuous critical thinking of the decision-maker or person interacting with the system.

So-called *tools for thought* (TfT) aim to promote critical thinking and reflection. Reflection helps to examine information and scrutinise assumptions, and thus has the potential to improve reasoning and judgement [11, 19, 22]. One way to promote reflection is through questions. Currently, the use of question in the context

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI2026 Workshop on Tools for Thought, Barcelona, Spain

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

of human-AI decision-making is relatively underexplored. Related work has been done by Reicherts et al. [20] who implemented a chatbot to probe people’s thinking with context-dependent questions. Another study found that phrasing causal explanations as questions can help people to better assess the validity of statements [2]. Moreover, in the field of education, chatbots are proposed that utilise the Socratic questioning technique, a common method in education [4, 14, 16]. Despite these approaches, many AI systems for decision-making focus on providing answers or advice in the form of predictions, recommendations and explanations.

Explanations help to understand how the system arrived at an outcome. In some cases, however, explanations can increase over-reliance, for example if the AI recommendation is incorrect [1, 23]. Furthermore, some operators may not be concerned with looking *inside* the “black-box” and trying to understand how the system works [24], but would rather like support in the broader decision-making context [13]. Explanations represent a backward-reasoning from the end result back to the input data [25]. Instead, decision-makers should be promoted in a forward-looking manner [7, 25]. This includes helping the decision-maker to develop their own line of reasoning, evaluate different options [15], and consider the consequences of the decision.

Questions can promote forward-looking reasoning, as they do not provide an answer and allow the person to think for themselves and form their own judgement, i.e., decision self-efficacy. Furthermore, questions can help to make implicit assumptions more explicit, thereby increasing awareness of the reasons behind a particular decision. The ability to recognise and articulate reasons makes it possible, on the one hand, to justify actions retrospectively and thus to be accountable (backward-looking), and more importantly, to direct actions in such a way that a desired state is achieved, thereby taking forward-looking responsibility [7].

Against this background, we therefore propose using questions to support the decision-making process. In the following we will briefly outline the various steps we have taken in this regard, aiming to contribute to the discussion on the design, development, and evaluation of tools for thought:

- A taxonomy for identifying relevant elements in human-AI decision-making, to which questions can relate (section 2).
- A prototype in the medical field (section 3.1) with preliminary feedback from clinicians (section 3.2).
- A method for generating data-driven questions using a large language model (section 4).
- A new proposed self-report scale for measuring cognitive engagement in human-AI decision-making (section 5).

The question taxonomy and the scale for measuring cognitive engagement can be found in the references provided [6, 7]. The prototype including clinicians’ experiences and the method for generating questions with LLMs are currently work in progress [5, 8].

2 A Taxonomy for Identifying and Formulating Relevant Questions

In order to identify and formulate relevant questions that stimulate critical reflection, we have created a taxonomy of elements in human-AI decision-making to which questions can relate to [7].

To this end, we adapted a taxonomy of Socratic questions [18] and mapped it to the context of human-AI decision-making using a question bank for explainable AI [13]. We also used Bloom’s taxonomy to argue that questions should address higher-order cognitive processes.

We identify 10 dimensions or elements that questions can relate to, such as model behaviour and decision boundaries (e.g., *Would you suggest the same treatment if the patient were 5 years older?*), assumptions or cognitive biases of the decision-maker (e.g., *How does the machine recommendation compare to your assumptions?*), or relevance of data points (e.g., *Is factor x relevant to focus on?*). These questions are aimed to help the decision-maker to reconsider the information relating to the patient’s case, understand how the DSS works, and reflect on the decision at hand.

While creating the taxonomy, we focused on decision-makers who operate a DSS. Hence, the taxonomy aims to help formulate questions that promote the cognitive engagement among decision-makers. Nevertheless, other stakeholders like developers could also benefit from the taxonomy, by questioning assumptions or the appropriateness of a data set. Furthermore, the taxonomy could be used as a means of promoting AI literacy, for example by helping people to remain more critical of AI-generated information.

3 Questions in Clinical Decision-Making

3.1 Prototype Design

We implemented a web-based prototype for a medical decision-task, namely the treatment of chronic low back pain, to assess the feasibility of a system that generates questions and to evaluate its perceived usefulness in the decision-making process [8]. For this, we replicated a decision-support system that is used in clinical practice for several years to help clinicians make decisions about treatment options. Importantly, we added the functionality of outputting five different questions based on patient information, the DSS behaviour, and its predictions. Accordingly, the interface shows a question alongside the treatment prognoses (Fig. 1).

To identify possible questions our system could output, we drew on the question taxonomy described above. We inspected the model behaviour and its decision boundaries through feature perturbation in order to extract information that can be questioned. One question, for example, addresses a hypothetical scenario, similar to counterfactual explanations, and asks about the feasibility of changing certain factors. For this, alternative predictions are calculated by considering all possible feature combinations. The operator then selects, as shown in Fig. 1, the desired treatment option (first drop-down) and whether the predicted effectiveness for that treatment option should be increased or decreased (second drop-down). Then, the alternative prediction is selected which, compared to the current prediction, shows the smallest change in input features, i.e., patient information, and the greatest effect on the outcome, i.e., treatment prediction. In addition, we also implemented a general question that is not informed by data. The five questions are listed in Table 1.

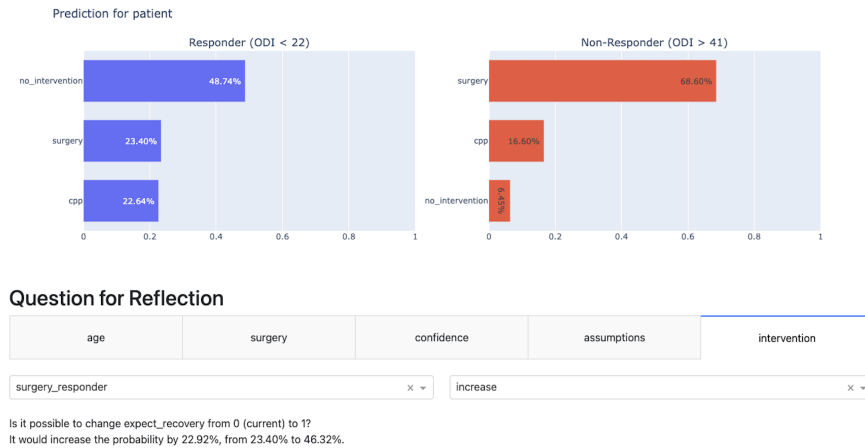


Figure 1: The interface of our prototype: The bar charts show predictions of the effectiveness of three possible treatment options, divided into responder (success) and non-responder (failure) categories. At the bottom, reflective questions are displayed in plain text. The current tab shows the question about the possibility to change an input feature to make the effectiveness of a treatment option more likely (minimum change in input with the maximum effect in outcome). This counterfactual thinking can provide insights into the workings of the DSS. In addition, it could provide an opportunity to first consider other intervention options, such as therapy, and only then consider other treatments, such as surgery.

Table 1: The five questions that our prototype outputs alongside the treatment recommendation. We used our question taxonomy to identify relevant questions.

Taxonomy ID	Question
Q10	Is the patient’s age of 47 years relevant to consider in this case? The patient of 47 years is approaching the red-flag threshold which is 50 years. Age correlates with the effect of surgery.
Q1	When was the specified surgery performed and at which location of the spine? Previous surgeries reduce the effect of surgery by 15% - 25%.
Q6	How confident are you about your decision? The confidence of the prediction for the most effective treatment (surgery 59.92%) is at 42.58%.
Q6	Does the prediction change your initial judgement? If so, why?
Q9	Is it possible to change the patient’s <i>expected recovery</i> from ‘no’ (current) to ‘yes’? It would increase the expected effectiveness of surgery by 22.92%, from 23.40% to 46.32%.

3.2 Preliminary Findings: Perceptions of Clinicians

We presented our prototype to clinical spine experts (n=6) and asked them about their perceptions in semi-structured, in-situ interviews [8]. We recorded and transcribed the interviews and analysed the data using thematic analysis to gain initial insights into our two research questions:

- **RQ1.** How do clinicians perceive questions during decision-making?
- **RQ2.** What makes questions effective from a clinician’s perspective?

The participants reported a high familiarity with our prototype, since the functionality of our prototype and the visual representation of the treatment predictions are based on the DSS they use in clinical practice. Regarding *RQ1*, we identified five themes, namely 1) questions can provide insights into how the DSS works, 2) questions can function as reminders to check information, 3) questions can prompt discussions with patients, 4) questions can help to consider alternatives to the most probable option, and 5) questions are perceived differently depending on the person.

The general sentiment in the interviews was that the questions presented in our prototype would not be particularly helpful to clinicians in reflecting on their decision-making. This is because, as the participants noted, certain factors (i.e., features) referred to in the questions were either part of their standard reasoning (e.g., previous surgeries), or were considered less crucial (e.g., patient’s age). Nevertheless, the clinicians also reported that questions could serve as reminders to check or consider certain information, particularly when one is less focused, such as at the end of the day, or that questions could help novice clinicians better understand how the DSS works, thereby preventing them from simply relying on the DSS prediction. For example, although the DSS takes into account the patient’s previous surgeries, the length of time since the operation is significant. A previous surgery that has a great impact on the DSS prediction may be considered negligible if it was performed 20 years ago.

Although we have found that participants perceive the usefulness of a particular question differently, we identified five characteristics that make questions more effective (*RQ2*), namely: 1) Questions

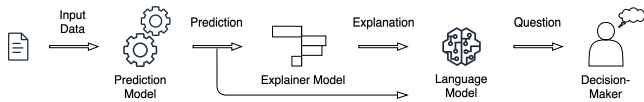


Figure 2: A flowchart illustrating our question generation system. Based on input data, such as patient information, a decision-support system computes a prediction. An explanation is then generated for this prediction in the form of feature contribution (LIME). Finally, the explanation and the prediction are passed to a language model to formulate a question. The generated question helps the decision-maker reflect on the prediction and decision at hand.

must fit the context, 2) questions must address factors that are not part of standard reasoning, 3) questions must be answerable and actionable, 4) the timing of questions matters, and 5) questions should not be too demanding.

4 Generating Questions with a Language Model

The initial questions in our current prototype are predefined. To make the generation of questions more flexible and generalisable, we propose an approach that utilises a large language model [5]. More specifically, we take the DSS prediction, generate an explanation in the form of feature contribution through LIME, and prompt a local language model with this information (Fig. 2). Our prompts are designed to generate questions that align with our question taxonomy.

As such, the feature contribution, i.e., LIME explanation, allows to generate questions that relate, for example, to the most important feature and its relevance (taxonomy ID Q2), or to the features with a negative contribution, asking why the option should be considered despite the factors that speak against it (taxonomy ID Q4). An example prompt could be: *“The most effective treatment is likely to be [prediction]. The features for this prediction are [positive_features], while the features against this prediction are [negative_features]. Formulate a question that stimulates the decision-maker to reflect on the prediction, asking why they would go for this decision, despite the reasons against it. Keep the question concise.”* A resulting example question is: *“Why would you prioritize conservative care despite concerns about recovery expectation, prior surgery, and neuromuscular conditions?”*

With this approach, which is related to natural language explanations in the field of human-centred explainable AI, it is possible to translate technical explanations in the form feature contribution, which might be difficult to understand for decision-makers, into a question that promotes critical thinking.

We noticed, however, that a small amount of the LLM-generated questions contained variables, e.g., hemoglobin levels, that do not appear in our dataset of patient cases and are thus not listed in the explanation, i.e., the feature contributions. This should remind us that LLMs are statistical models that generate text based on patterns derived from the data they were trained on. To minimise the risk of generating irrelevant or incorrect questions, which in the worst case could be misleading, it would be possible to implement

a RAG (retrieval-augmented generation) approach in order to add contextual or domain-specific knowledge.

5 A Scale for Measuring Cognitive Engagement

Current evaluation methods of decision-support systems or other interventions to mitigate overreliance focus on the decision accuracy, i.e., whether the correct decision was made. We argue, however, that evaluation of tools for thought must take into account the decision-making process. As such, it should be considered whether the TfT helped to consider alternatives, or to understand the available options. We therefore propose possible items for a new self-report scale designed to measure cognitive engagement in human-AI decision-making, and thus the effectiveness of tools for thought such as our question-generation system [6]. Accordingly, example items of a scale could include, *“The system (TfT) helped me to be aware of my preferences/assumptions”*, or *“The system (TfT) helped me to compare and contrast different options”*.

We have derived and adapted potential items from existing scales used in education to assess how students use technology to perform cognitive learning activities, as well as to measure associated factors to cognitive engagement, such as decision self-efficacy. We have not used this scale in our exploratory evaluation of our prototype (section 3), as we developed the potential scale in parallel with the first feedback sessions with the clinicians. Future work is necessary to develop a validated scale.

In addition to assessing the impact of generative AI on critical thinking and reflection, such a scale could serve as a guide for the design and development of future tools for thought aimed at promoting cognitive engagement.

6 Concluding Remarks

Some open questions remain for future work. Namely, how to formulate relevant questions in more open-ended human-AI interactions. Generating relevant questions might be more difficult once there is no decision-support system with explainable AI, i.e., underlying logic/structure how input data connects to an outcome.

Moreover, as our exploratory study suggests, questions can be perceived differently by different people. Therefore it will be important for various tools for thought, such as our question-generation system, to adapt to individual decision-makers. People differ in their decision-making styles, their need for cognition, and their risk attitudes – all factors that influence decision-making. Moreover, external factors like time constraints also play a role. Future work would need to investigate how these aspects can be taken into account in order to generate relevant questions. Our proposed approach of using a language model could serve as a first step towards adapting and personalising questions.

Since reflection has been shown to improve decision-making, it seems promising to utilise generative AI in this regard to promote cognitive engagement, forward reasoning, and decision self-efficacy.

Acknowledgments

This research is funded by the Donders Centre for Cognition. We thank two anonymous reviewers for their constructive feedback.

References

- [1] 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*. ACM, Yokohama Japan, 1–16. doi:10.1145/3411764.3445717
- [2] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. 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, 1–13. doi:10.1145/3544548.3580672
- [3] Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi, Roman KloECKner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto dos Santos. 2023. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology* 307, 4 (May 2023), e222176. doi:10.1148/radiol.222176
- [4] Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing Critical Thinking in Education by Means of a Socratic Chatbot. arXiv:2409.05511 [cs]
- [5] Simon W.S. Fischer, Linus Holmberg, Serge Thill, and Hanna Schraffenberger. 2026. From Explanations to Questions: Adding Friction and Promoting Cognitive Engagement in Human-AI Decision-Making. Manuscript submitted for publication..
- [6] Simon W.S. Fischer and Hanna Schraffenberger. 2025. How to Measure Cognitive Engagement in Machine-Assisted Decision-Making?. In *Proceedings of the Workshops at the Fourth International Conference on Hybrid Human-Artificial Intelligence Co-Located with the Fourth International Conference on Hybrid Human-Artificial Intelligence (HHAI 2025)*, Vol. 4074. CEUR Workshop Proceedings, Pisa, Italy, 98–109.
- [7] 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 (October 2025), 940–954. doi:10.1609/aies.v8i1.36602
- [8] Simon W.S. Fischer, Hanna Schraffenberger, Miranda van Hooff, Serge Thill, and Pim Haselager. 2026. Questions Can Provide Insights and Function as Reminders: Clinicians' Perceptions of A Decision-Support System That Presents Data-Driven Questions. Manuscript submitted for publication..
- [9] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection. *Translational Psychiatry* 11, 1 (Feb. 2021), 108. doi:10.1038/s41398-021-01224-x
- [10] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3581196
- [11] Zohreh Khoshgoftar and Maasoumeh Barkhordari-Sharifabad. 2023. Medical Students' Reflective Capacity and Its Role in Their Critical Thinking Disposition. *BMC Medical Education* 23, 1 (March 2023), 198. doi:10.1186/s12909-023-04163-x
- [12] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. ChatGPT's Inconsistent Moral Advice Influences Users' Judgment. *Scientific Reports* 13, 1 (April 2023), 4569. doi:10.1038/s41598-023-31341-0
- [13] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, Honolulu HI USA, 1–15. doi:10.1145/3313831.3376590
- [14] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, Vol. 37. Curran Associates, Inc., 85693–85721.
- [15] Tim Miller. 2023. 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, 333–342. doi:10.1145/3593013.3594001
- [16] Hadeel Naeem. 2025. Teaching Skills and Intellectual Virtues with Generative AI. *Episteme* (November 2025), 1–18. doi:10.1017/epi.2025.10089
- [17] Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. 2024. *Appropriate Reliance on Generative AI: Research Synthesis*. Technical Report MSR-TR-2024-7. Microsoft.
- [18] Richard Paul and Linda Elder. 2019. *The Thinker's Guide to Socratic Questioning*. Rowman & Littlefield Publishers, Blue Ridge Summit.
- [19] Shivesh Prakash, Ruth M. Sladek, and Lambert Schuwirth. 2019. Interventions to Improve Diagnostic Decision Making: A Systematic Review and Meta-Analysis on Reflective Strategies. *Medical Teacher* 41, 5 (2019), 517–524. doi:10.1080/0142159X.2018.1497786
- [20] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces*. ACM, Glasgow United Kingdom, 1–10. doi:10.1145/3543829.3543832
- [21] Mrinank Sharma, Miles McCain, Raymond Douglas, and David Duvenaud. 2026. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. arXiv:2601.19062 [cs] doi:10.48550/arXiv.2601.19062
- [22] Carolina Walger, Karina De Dea Roglio, and Gustavo Abib. 2016. HR Managers' Decision-Making Processes: A "Reflective Practice" Analysis. *Management Research Review* 39, 6 (June 2016), 655–671. doi:10.1108/MRR-11-2014-0250
- [23] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. doi:10.1145/3397481.3450650
- [24] Carlos Zednik. 2019. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34, 2 (Dec. 2019), 265–288. doi:10.1007/s13347-019-00382-7
- [25] Zelun Tony Zhang, Sebastian S. Feger, Lucas Dullenkopf, Rulu Liao, Lou Süßlin, Yuanqing Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (November 2024), 1–32. doi:10.1145/3687024

Received March 27, 2026