Using Imperfect AI to Support High-Stakes Decisions

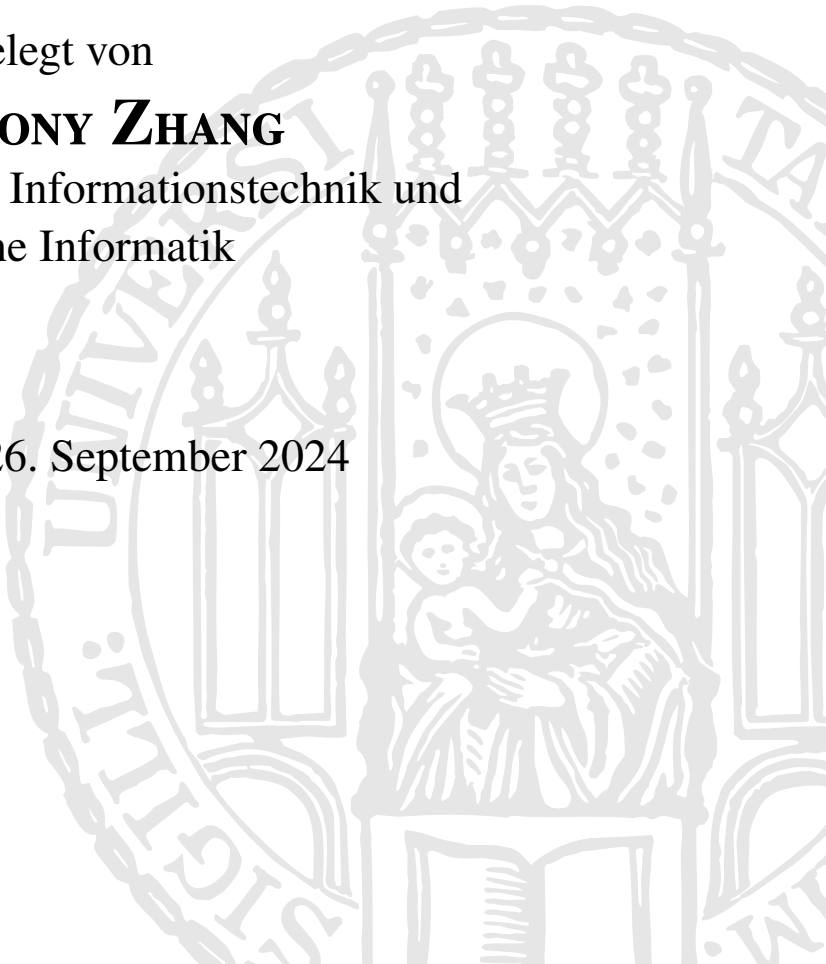
Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

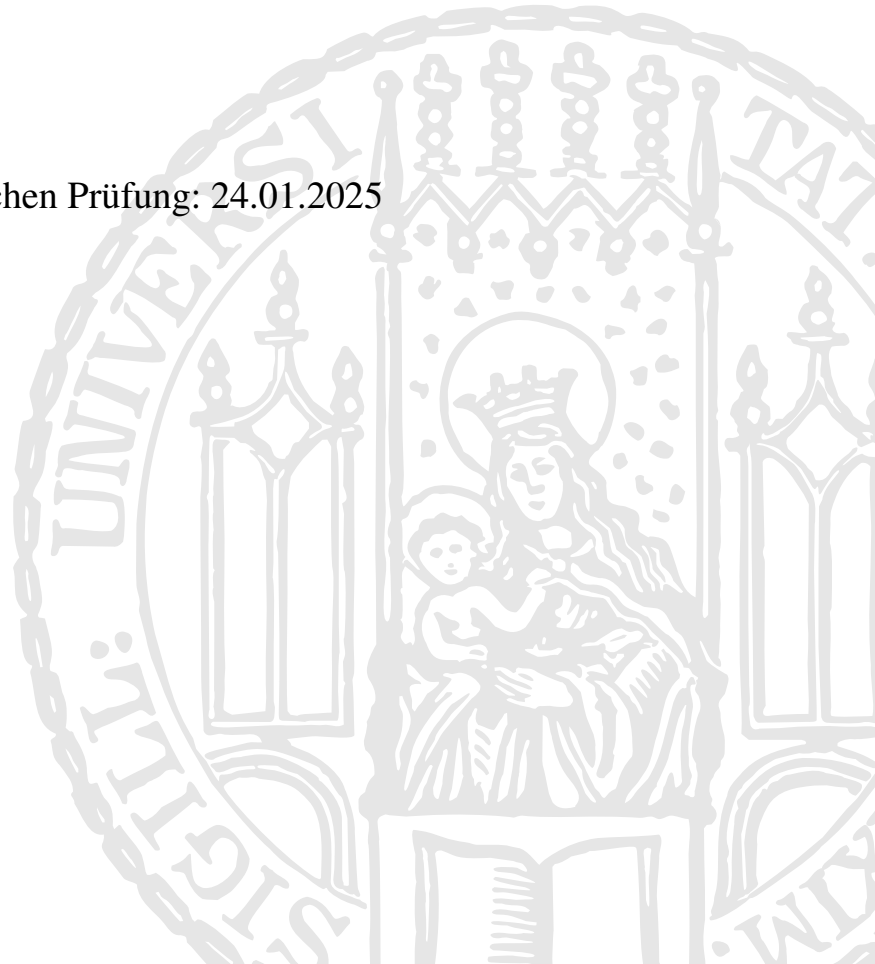
Zelun Tony Zhang
M.Sc. Elektrotechnik, Informationstechnik und
Technische Informatik

München, den 26. September 2024



Erstgutachter:	Prof. Dr. Andreas Butz
Zweitgutachter:	Prof. Dr. Daniel Buschek
Drittgutachter:	Prof. Krzysztof Gajos

Tag der mündlichen Prüfung: 24.01.2025



ABSTRACT

Given the impressive technological advances of recent years, artificial intelligence (AI) is anticipated to enhance human decision-making across various fields, including challenging, high-stakes ones like healthcare, finance, or aviation. However, recent research shows that supporting human decision makers with AI is far from straight-forward.

Often, AI decision support is designed to be recommendation-centric, i.e., the AI gives a closed, end-to-end decision recommendation. Since the human is not involved in generating the recommendation, they often cannot work effectively with it, particularly when the AI is wrong. Incorrect recommendations can severely disrupt the human decision maker's workflow or, if unnoticed, lead them to adopt the wrong recommendation (overreliance), which is especially problematic for high-stakes decisions. A common approach to address this is to explain the AI recommendation, but results have been mixed so far. This thesis advances the understanding of how AI—despite its imperfection—can effectively support human decision-making in high-stakes domains, by addressing two major issues:

1) Through experiments on simple, controllable decision tasks with lay users, the thesis contributes a *better understanding of the limitations of recommendation-centric support* for high-stakes decisions. The experiment results reveal that humans are particularly prone to overrely on AI recommendations in challenging decisions and after repeated interactions with the AI. Common feature-based explanations could not effectively mitigate overreliance.

2) Through more qualitative studies with domain experts, the thesis explores *alternative approaches to AI decision support* to overcome the identified limitations. As explored for the use case of diversions in aviation, one promising approach is to design AI to support the process leading up to the final decision, rather than jumping straight to the end of the process with an end-to-end recommendation. The AI support must be granular and flexible enough for the decision maker to use it according to their own momentary intention. Instead of "opening the black box", explanations should help users understand how the AI output fits their intention. Finally, recommendations can be beneficial, but should be introduced toward the end of the decision-making process.

Overall, this thesis contributes to a more holistic view of the design space of AI-driven decision support that goes beyond recommendations and explanations. Discovering these alternative design opportunities requires a thorough understanding of users' decision-making processes and workflows, highlighting that the classic endeavor of human-centered design to start from human needs remains crucial also in the age of advanced AI technology.

ZUSAMMENFASSUNG

Angesichts der beeindruckenden technologischen Fortschritte der letzten Jahre wird davon ausgegangen, dass künstliche Intelligenz (KI) die menschliche Entscheidungsfindung in den unterschiedlichsten Bereichen verbessern wird, darunter auch in anspruchsvollen und riskanten Bereichen wie dem Gesundheitswesen, dem Finanzwesen oder der Luftfahrt. Jüngste Untersuchungen zeigen jedoch, dass die Unterstützung menschlicher Entscheidungsträger durch KI alles andere als einfach ist.

Häufig ist KI-Entscheidungsunterstützung empfehlungsorientiert, d. h., die KI gibt eine geschlossene Entscheidungsempfehlung ab. Da der Mensch nicht an der Erarbeitung der Empfehlung beteiligt ist, kann er oft nicht effektiv damit arbeiten, insbesondere wenn die KI falsch liegt. Fehlerhafte Empfehlungen können die Arbeit des menschlichen Entscheidungsträgers empfindlich stören oder, wenn sie unbemerkt bleiben, dazu führen, dass er die falsche Empfehlung annimmt (Overreliance), was besonders bei risikoreichen Entscheidungen problematisch ist. Ein gängiger Lösungsansatz besteht darin, die KI-Empfehlung zu erklären, doch die Ergebnisse sind bisher durchwachsen. Diese Dissertation erweitert das Verständnis dafür, wie KI trotz ihrer Unvollkommenheit die menschliche Entscheidungsfindung in risikoreichen Domänen effektiv unterstützen kann, indem sie zwei Hauptthemen behandelt:

1) Durch Experimente an simplen, kontrollierbaren Entscheidungsaufgaben mit Laien trägt die Dissertation zu *einem besseren Verständnis der Limitierungen von empfehlungszentrierter Unterstützung* für risikoreiche Entscheidungen bei. Die Ergebnisse zeigen, dass Menschen besonders bei schwierigen Entscheidungen und nach wiederholten Interaktionen mit der KI dazu neigen, sich übermäßig auf KI-Empfehlungen zu verlassen. Übliche merkmalbasierte KI-Erklärungen konnten das übermäßige Verlassen nicht wirksam eindämmen.

2) Durch qualitative Studien mit Domänenexperten werden in der Dissertation *alternative Ansätze zur KI-Entscheidungsunterstützung* untersucht, um die identifizierten Einschränkungen zu überwinden. Wie für den Anwendungsfall der Flugumleitungen in der Luftfahrt untersucht, besteht ein vielversprechender Ansatz darin, mit KI den Prozess bis zur endgültigen Entscheidung zu unterstützen, anstatt direkt ans Ende des Prozesses zu springen und eine abgeschlossene Empfehlung abzugeben. Die KI-Unterstützung muss so granular und flexibel sein, dass der Entscheidungsträger sie je nach seiner momentanen Absicht einsetzen kann. Anstatt die "Blackbox zu öffnen", sollten KI-Erklärungen den Nutzern helfen zu verstehen, wie die KI-Ausgabe zu ihren Absichten passt. Zudem können Empfehlungen zwar nützlich sein, sollten aber erst am Ende des Entscheidungsprozesses eingeführt werden.

Insgesamt trägt diese Dissertation zu einer ganzheitlicheren Sicht auf den Gestaltungsraum für KI-Entscheidungsunterstützung bei, die über KI-Empfehlungen und -Erklärungen hinausgeht. Um diese alternativen Gestaltungsmöglichkeiten zu identifizieren, ist ein gründliches Verständnis der Entscheidungsprozesse und Arbeitsabläufe der Nutzer erforderlich. Dies zeigt, dass das klassische Bestreben der menschenzentrierten Gestaltung, von den menschlichen Bedürfnissen auszugehen, auch im Zeitalter fortschrittlicher KI-Technologie entscheidend bleibt.

ACKNOWLEDGMENTS

This thesis has been significantly shaped by the people I have shared my PhD journey with, and I am grateful for every single person I have met along the way. First of all, thank you to **Heinrich** for your supervision, for encouraging me in what I was doing, and for warning me not to reinvent the wheel. I wish you could have seen how it all turned out. Thank you also to **Andreas** for willingly taking over the supervision, which gave me much-needed security. Thank you for encouraging me not to overthink everything and for pushing me over the finish line to get these pages written during an intense time in my life. Thank you to **Daniel** and **Krzysztof** for reviewing my thesis and for giving supportive feedback for my work.

Thank you also to my colleagues at fortiss, particularly to my Human-centered Engineering teammates. Thank you to **Yuanting** for trusting me and for giving me the freedom to explore. Thank you to **Zhiwei**, **Sören**, **Cara**, and **Mariam** for many fun conversations and for the great collaboration during proposal writing and research projects. Thank you especially to Cara for pushing the KIEZ project forward with your ideas and to Mariam for bringing in your research experience. I have learnt a lot from working with both of you. I also want to say a big thank you to **Swen** and **Markus** from the fortiss IT team for always being incredibly helpful with my various requests and equipment issues.

Thank you also to the entire LMU MIMUC team, you have always welcomed me warmly and I have benefited immensely from hearing about your research and discussing my work with you. Thank you especially to **Sebastian** for enthusiastically joining me on my aviation research project. Your support took the project to the next level. Special thanks also to **Thomas** and **Rainer** for answering all my questions about processes at the group, to **Amy** for welcoming me in your office, and to **Changkun** and **Florian** (Lang) for sharing a great time at IUI 2023 in Sydney. Thank you also to **Hai** and **Florian** (Lehmann) for great times as "fellow externals" during IDCs and conferences.

Thank you also to my students, **Jessica**, **Yuliia**, **Jenny**, **Amina**, **Sven**, **Felicitas**, **Marc**, **Lou**, and **Rulu**. Working with you has been one of the most enjoyable parts of my PhD, and each of your works has contributed to my research in one way or another. I hope it has been a fruitful learning experience for you—it certainly has been for me!

Lastly, thank you to my two girls. Thank you, **Clara**, for bringing so much sunshine into our lives. One broad smile or joyful laugh from you and the troubles of the day are forgotten. Thank you also for reliably taking your naps, so that I could write this thesis with you sleeping peacefully by my side. But the biggest thanks go to you, **Qianyu**. This PhD would not have been possible without your unwavering support, especially when paper deadlines were approaching. I feel so blessed to have you by my side, and I look forward to the future that lies ahead of us.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Statement

Fueled by the rapid technological advances in recent years, artificial intelligence (AI) has become ubiquitous, permeating more and more aspects of daily life and work. One popular application of AI is to support human decision-making across various domains [50], often in high-stakes ones such as human resources [85], finance [15], or healthcare [103]. Common terms for this field of research include *human-AI decision-making* or *AI-assisted decision-making*. Throughout the thesis, I will use the latter term to emphasize the human-centered perspective that informs all of the work in this thesis: that it is humans who make the decisions and who are responsible for their decisions, with AI only providing assistance, rather than being an equal partner.

A common assumption is that AI will augment human decision-making, given the complementary strengths of humans and machines: While humans are better at understanding context and human values, machines can process far more data [1, 42]. However, AI is imperfect, and some argue that it will always be [93]. The question is therefore how to benefit from AI without AI errors leading to grave consequences in high-stakes applications. A frequently proposed solution is the *human-in-the-loop* [1, 21, 50, 84], i.e., relying on the human to override the AI when necessary. But recent empirical results show that this is far from trivial, since humans frequently override or accept AI recommendations inappropriately, as discussed in more detail in Section 2.1. As a result, the vision of AI augmenting human decision-making, especially in high-stakes applications, has proven difficult to realize.

The aim of this thesis is to contribute toward the realization of this vision through thoughtful human-AI interaction design. The thesis is grounded in three interrelated observations:

1. Systems for AI-assisted decision-making are usually recommendation-centric, i.e., the AI system provides end-to-end decision recommendations, as I will elaborate further in Section 2.2.

2. This recommendation-centric support paradigm may be particularly prone to AI mistakes, as it depends on the ability of the human decision maker to recognize a bad recommendation.

3. AI-assisted decision-making does not have to be recommendation-centric. There are often many more opportunities to support decision-making with AI than end-to-end recommendations.

Motivated by these observations, I pursued two main lines of research: (1) understanding the limitations of recommendation-centric decision support, and (2) how alternative support paradigms might overcome these limitations. The results demonstrate the importance of understanding human decision-making processes and the potential of approaches beyond the common recommendation-centric paradigm.

1.2 Contributing Publications

This thesis is a cumulative dissertation that connects and summarizes the contributing publications listed below and puts them into the context of related work. Throughout the thesis, I reference these publications with a prefixed "P" (e.g., [P1]). An overview of my own contributions to these publications is given in Table 1.1. More details including clarifications of my co-authors' contributions will be given in the respective sections of Chapter 3. This is meant to improve the reading flow, as the context of each publication should help to better understand each author's contribution.

Table 1.1: Overview of own contributions to the publications. *Main contributor* means that these contributions were mostly or entirely mine. *Co-contributor* means that these contributions were shared with my co-authors. Note that a lack of entries under *co-contributor* does not mean a lack of contribution by my co-authors, but only that there was no shared contribution between me and my co-authors. More details about my own as well as my co-authors' contributions are given in the respective sections in Chapter 3.

Publication	Main contributor	Co-contributor
Pilot Attitudes Toward AI in the Cockpit [P1]	study idea, design, and execution; data analysis; writing	
Forward Reasoning Decision Support [P2]	idea for position paper; writing	
Resilience Through Appropriation [P3]	writing	study design and execution; prototype design; data analysis
Is Overreliance on AI Provoked by Study Design? [P4]	study idea; data analysis; writing	study design
You Can Only Verify When You Know the Answer [P5]	study idea; data analysis; writing	study design
Beyond Recommendations [P6]	study idea, design, and execution; system design and implementation; data analysis; writing; framework for process-oriented support	focus group execution
Effect of Mental Workload and Explanations on Appropriate AI Reliance [P7]	study idea; implementation of AI model	study design; data analysis; writing

[P1] Zelun Tony Zhang, Yuanting Liu, and Heinrich Husmann. Pilot attitudes toward AI in the cockpit: implications for design. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pp. 33:1–33:6, Magdeburg, Germany, September 2021. IEEE. DOI: 10.1109/ICHMS53169.2021.9582448.

[P2] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. Forward reasoning decision support: toward a more complete view of the human-AI interaction design space. In *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, pp. 18:1–18:5, Bolzano, Italy, July 2021. ACM. DOI: 10.1145/3464385.3464696.

[P3] Zelun Tony Zhang, Cara Storath, Yuanting Liu, and Andreas Butz. Resilience through appropriation: Pilots' view on complex decision support. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pp. 397–409, Sydney, NSW, Australia, March 2023. ACM. DOI: 10.1145/3581641.3584056.

[P4] Zelun Tony Zhang, Sven Tong, Yuanting Liu, and Andreas Butz. Is overreliance on AI provoked by study design? In José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2023*, pp. 49–58, York, UK, August 2023. Springer. DOI: 10.1007/978-3-031-42286-7_3.

[P5] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. You can only verify when you know the answer: Feature-based explanations reduce overreliance on AI for easy decisions, but not for hard ones. In *Proceedings of Mensch und Computer 2024*, MuC '24, pp. 156–170, Karlsruhe, Germany, September 2024. ACM. DOI: 10.1145/3670653.3670660.

[P6] Zelun Tony Zhang, Sebastian S. Feger, Lucas Dullenkopf, Rulu Liao, Lou Süsslin, Yuanting Liu, and Andreas Butz. Beyond recommendations: From backward to forward AI support of pilots' decision-making process. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):485:1–485:32, November 2024. DOI: 10.1145/3687024.

[P7] Zelun Tony Zhang, Seniha Ketenci Argın, Mustafa Baha Bilen, Doğan Urgun, Sencer Melih Deniz, Yuanting Liu, and Mariam Hassib. Measuring the effect of mental workload and explanations on appropriate AI reliance using EEG. *Behaviour & Information Technology*, pp. 1–19, November 2024. DOI: 10.1080/0144929X.2024.2431055. Advance online publication.

Publication [P5] received a *Best Paper Award*.

Chapter 2

Background and Definitions

2.1 Appropriate Reliance

This thesis focuses on AI-assisted decision-making in high-stakes applications, where bad decisions can have highly negative consequences. It is therefore crucial that decision makers are not misled into a bad decision by an inevitable AI error. Conversely, it is desirable that the decision maker does not erroneously overrule a good AI recommendation. This notion is captured by the term *appropriate reliance*, which has already been studied for decades in the context of traditional automation [51], and has in recent years also gained much attention in AI-assisted decision-making research [75].

Appropriate reliance is usually defined and measured through decision outcomes on individual decision instances [32]. The setup is typically such that the AI makes a recommendation to the human, who has to make the final decision. If the human accepts a correct AI recommendation or rejects a wrong one, the reliance is said to be appropriate. Conversely, reliance is said to be inappropriate if the human either accepts a wrong AI recommendation (*overreliance*), or rejects a correct recommendation (*underreliance*). This definition of appropriate reliance is straightforward to operationalize by recording participants' decisions conditioned on the correctness of the AI recommendation.

While this outcome-based definition of appropriate reliance is very common (e.g., [11, 38, 56, 72, 75, 88, 92, 97], it has its shortcomings, as highlighted by Fok and Weld [32]. For one, relying on AI can be considered an appropriate strategy even if the AI ends up being wrong, which could be the case for instance when a typically highly reliable AI makes a rare, unexpected mistake. The reverse of this argument also holds. Second, the outcome-based definition does not properly consider non-deterministic elements in decision outcomes. For example, two loan applicants may have the same profiles, resulting in the same AI recommendations, but one may default while the other does not. Fok and Weld therefore propose a "strategy-graded" definition for appropriate reliance [32], where the appropriateness does not depend on the outcome of a decision, but on whether the AI is expected to outperform the

human for a given decision. However, it is often unclear how to operationalize this definition in experiments.

Another limitation of the outcome-based definition is that it does not allow to discriminate different reasons why people do not make optimal use of AI recommendations. To address this limitation, Guo et al. [37] propose a decision-theoretic framework that differentiates between an overall reliance level and the recognition whether the human or the AI is better for a specific decision instance. The framework relies on strong assumptions about people's decision-making process with AI, which are yet to be validated and refined in future work.

In this thesis, I resorted to the established outcome-based definition of appropriate reliance. While I recognize its limitations, work on alternative definitions is still in its infancy and represents an important area for future research. I addressed the limitations of the outcome-based definition through the experiment designs, e.g., by choosing deterministic decision tasks or by carefully selecting the decision instances to include in the experiments. I further focused primarily on overreliance. While underreliance is also undesirable, as it means missed opportunities for potentially critical improvements, overreliance means the AI introduced a change for the worse. Overreliance is arguably also the greater concern both in public debates about AI as well as in research on AI-assisted decision-making [11, 47, 88, 89].

2.2 Recommendation-Centric Decision Support

As previously stated in Section 1.1, the most common approach to AI-assisted decision-making is what I term *recommendation-centric decision support*. By this I refer to systems where the primary functionality is to provide end-to-end decision recommendations, which the human decision maker can either accept or reject. For example, a decision support tool in healthcare could suggest a likely diagnosis to physicians directly from the patient's information. However, numerous studies demonstrate that humans often rely inappropriately on AI recommendations, with both underreliance [17, 19, 22, 69] and—more frequently—overreliance [6, 11, 34, 49, 55, 67, 77] being an issue, which can even be the case for expert decision makers [10, 41].

To help people to rely on AI recommendations appropriately, the recommendations are often accompanied by explanations about how the AI model arrived at its output. The *explainable AI* (XAI) community has contributed numerous approaches to explain the output of an AI model [53], such as by revealing how individual input features contributed to the output (*feature-based explanations*), or by showing examples from the training set that are similar to the current input (*example-based explanations*). The rationale is that by "opening the black box" [36] of AI in this way, humans can judge whether the system is correct or not [74]. However, empirical results have been mixed so far. While explanations are beneficial in some studies [18, 88, 97], it seems more common that explanations have little effect on the appropriateness of reliance [7, 75, 92, 106], or even increase overreliance [6, 10, 41, 49, 77].

Besides explanations, several other approaches have been proposed to improve appropriate reliance on recommendation-centric decision support. One approach is to provide additional information about the AI model to decision makers, such as the model's test set accuracy [38, 71] or descriptions about the model performance [14]. One type of model information that has repeatedly been shown to be effective for calibrating people's reliance is model uncertainty about a given recommendation [6, 58, 68, 106]. The challenge for practical use is that uncertainty is very hard to calibrate, as AI can often be wrong with high confidence [4].

Another approach is the use of cognitive forcing functions [11], which aim to push decision makers to engage more deliberately with AI recommendations, e.g. by displaying recommendations only after the user has entered their own initial decision [11, 31] or by introducing a waiting time before the AI provides a recommendation [11, 65]. While cognitive forcing functions can mitigate overreliance, they are perceived negatively by users [11, 31, 65].

A further line of research is to optimize users' reliance behavior by adapting which of the above strategies to use, or even whether to show recommendations at all. Examples of proposed methods include reinforcement learning based on information such as the user's knowledge of a certain concept [12]; modeling the user's likelihood to make a correct decision in a given task instance [57]; or logically inferring adaptation strategies from simple assumptions about the user's reliance behavior [59]. Initial results on simple decision problems are promising, but the computational inferences introduce additional sources for mistakes. It is thus unclear how well these approaches translate into complex real-world, high-stakes applications.

All of the approaches mentioned above target the problem of inappropriate reliance, but remain focused on supporting human decision-making through end-to-end recommendations. However, besides inducing inappropriate reliance, recommendation-centric support also often proves less helpful to decision makers than expected, as revealed by a growing number of studies on AI-assisted decision-making under real-world conditions [8, 45, 98]. A recurring complaint of users is that recommendation-centric support does not integrate well into their existing work and decision-making processes [13, 46, 98]. Rather than a suggested end result, decision makers often wish for more support of the process leading up to the final decision [101, 105], which is completely ignored by recommendation-centric support. Such findings suggest that for many applications, a paradigm shift away from recommendation-centric support may be necessary for AI-assisted decision-making, although it is often unclear what alternative forms of AI support might look like, as concrete examples are rare.

2.3 Research Questions

While inappropriate reliance is a persistent problem in recommendation-centric support, there are studies showing that it is possible to calibrate human reliance on AI recommendations in some scenarios [97, 106]. For instance, Vasconcelos et al. [88] showed that explana-

tions can reduce overreliance when they lower the cost of verifying the AI recommendation. However, for most constellations, it is not as clear under which conditions people are able to rely appropriately on recommendation-centric support and when not. This motivates the first overarching research question of this thesis:

> ***RQ1:*** *What are the limitations of recommendation-centric support for high-stakes decisions?*

In particular, I focus on the most common form of recommendation-centric support, i.e., recommendations plus explanations. I further concentrate on overreliance, as explained in Section 2.1, and the conditions under which explanations fail to mitigate it.

As outlined in Section 2.2, many researchers attempt to address the issue of inappropriate reliance while holding on to the recommendation-centric paradigm of AI-assisted decision-making. Yet, studies on real-world decision contexts often reveal opportunities for AI support that is not recommendation-centric, even though examples of concrete system designs are rare. It is even less clear how such alternative support paradigms that do not rely on end-to-end decision recommendations would compare to recommendation-centric support. The second research question guiding the work in this thesis is therefore:

> ***RQ2:*** *How can AI support high-stakes decisions without being recommendation-centric to overcome the limitations of recommendation-centric support?*

By studying a use case from commercial aviation, my goal is to identify AI decision support designs that are useful to pilots, but not centered around recommendations, and compare them against the recommendation-centric approach.

Chapter 3

Imperfect AI for High-Stakes Decision Support

This chapter summarizes the contributing publications listed in Section 1.2 and relates them to the two guiding research questions laid out in Section 2.3. As these publications were the result of collaborations with my co-authors, I will use the scientific "*we*" throughout this chapter when describing the individual studies.

3.1 Limitations of Recommendation-Centric Support

> ***RQ1:*** *What are the limitations of recommendation-centric support for high-stakes decisions?*

I primarily investigated RQ1 through experiments on simple decision tasks (Sections 3.1.1 and 3.1.2), which is the most common way to empirically study AI-assisted decision-making [50]. While somewhat artificial, such simple decision tasks allow for controlled investigations of the human factors involved in AI-assisted decision-making. However, they also tend to conceal challenges that only become apparent in complex real-world use cases. I therefore also studied commercial aviation as a high-stakes domain to understand limitations arising in actual use (Section 3.1.3).

3.1.1 Increasing Overreliance Over Time

This section is based on the publication

***"Is overreliance on AI provoked by study design?"** [P4].*

The established outcome-based definition of appropriate reliance (Section 2.1) requires a study design where participants have to solve a series of decision tasks. In order to collect enough data, these task series are often quite long (see Table 3.1), long enough to suspect that many participants may become complacent and less engaged toward the end of the task series. At the same time, prior work has shown that overreliance in AI-assisted decision-making may be caused by a lack of cognitive engagement [11, 33]. We therefore asked whether observations of overreliance in empirical studies are provoked by the long task series used in many studies, which may not reflect how decisions are made in many real-world use cases.

Table 3.1: Examples of research finding overreliance in AI-assisted decision-making, including the decision tasks and the lengths of the task series participants had to complete in the respective studies. Adapted from [P4].

Publication	Study task	# Tasks
Bansal et al. [6]	Sentiment classification	50
	Law School Admission Test	20
Buçinca et al. [11]	Nutrition assessment	26
Green and Chen [34]	Recidivism risk assessment	40
	Loan risk assessment	40
Lai and Tan [49]	Deception detection	20
Liu et al. [55]	Recidivism prediction	20
	Profession classification	20
Schmidt and Biessmann [77]	Sentiment classification	50
Wang and Yin [92]	Recidivism prediction	32
	Forest cover prediction	32

To answer this question, we designed an online experiment with 47 participants based on a profession classification task that had been used in several other studies on AI-assisted decision-making [18, 55, 66, 78]. Participants had to read a short biography and decide with the help of recommendation-centric AI assistance which profession the described person has. The core idea was to design a study condition in which the 50 decision tasks were split into ten short sessions of five tasks each, which were sent to participants with a minimum break of one hour between two sessions. Participants were free to choose when they would complete the new session once they received the link to it. This condition (*multiple session*

group—MSG) was meant to induce less complacency among participants than the typical setup where all 50 tasks had to be completed in a single session (*single session group*—SSG).

The results showed that for both MSG and SSG, participants' agreement with the AI and their overreliance increased significantly as they progressed through the task series, while the decision time decreased significantly, indicating decreasing effort with ongoing time. This implies that typical study designs with long task series indeed favor the occurrence of overreliance. However, overreliance may increase over time even in settings like the MSG condition where people do not have to solve a large amount of decision tasks in a short time period. One reason could be that our MSG condition was not effective enough at reducing complacency. On the other hand, we ensured that participants could conveniently solve the tasks at their own schedule. Therefore, our findings potentially suggest a more fundamental limitation of recommendation-centric support, i.e., as people become more familiar with such a system, they become more likely to overrely on it, irrespective of whether they interact with it in long, tiring sessions, or in shorter sessions embedded into their daily context.

Author contributions: This publication is the result of the master thesis of my student Sven Tong. I provided the study idea, conducted the statistical analysis in the form found in the publication, and wrote the paper. Sven designed the study under my guidance, implemented the machine learning model, the explanations, and the study interface, and conducted the study. Yuanting Liu and Andreas Butz provided guidance and feedback and edited the final version of the submission.

3.1.2 Inappropriate Reliance in Difficult Decisions

This section is based on the following publications:

> *"**You can only verify when you know the answer: Feature-based explanations reduce overreliance on AI for easy decisions, but not for hard ones**" [P5],*
>
> *"**Measuring the effect of mental workload and explanations on appropriate AI reliance using EEG**" [P7].*

As shown by an increasing number of empirical studies, countless factors can influence whether people can rely appropriately on recommendation-centric decision support, such as users' personality traits [11, 77], their domain expertise [39, 92], model performance [61, 64, 102], or the type [89, 92] and design [97] of explanations. We were interested in the effect of task characteristics, in particular decision difficulty, as AI-assisted decision-making is often aimed at difficult decisions in high-stakes domains like healthcare [79] or finance [15], where overreliance is especially undesirable. At the same time, given their high stakes, the call for explainability is particularly prominent in these domains [2, 5]. We therefore conducted two experiments to investigate how decision difficulty affects the appropriateness of human reliance on AI, and how the effectiveness of explanations to mitigate overreliance depends on decision difficulty.

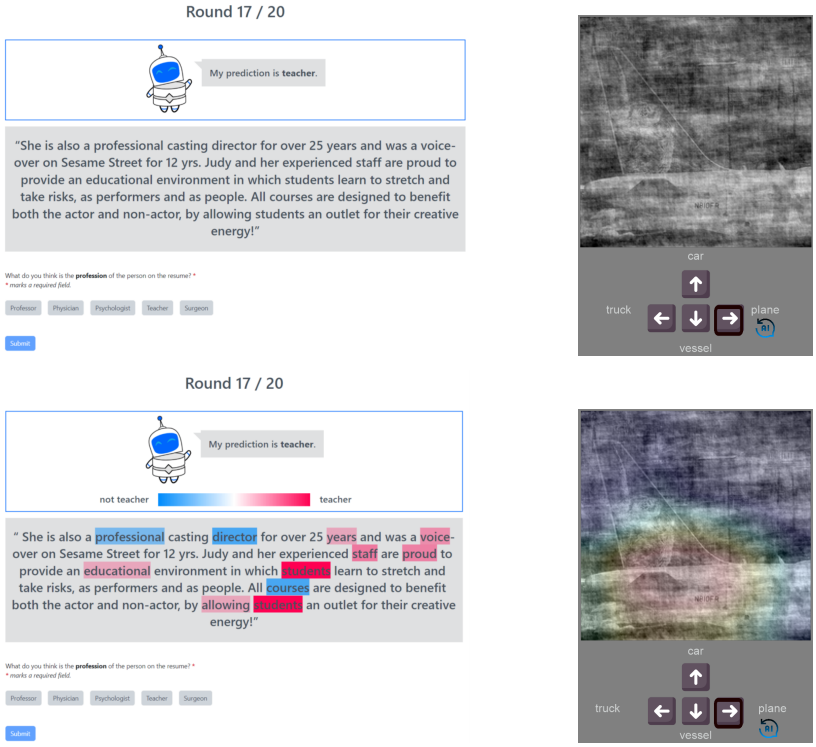


Figure 3.1: Task interfaces used in the decision difficulty experiments. Left: profession classification task used in [P5]. Right: image recognition task used in [P7]. Top: AI support without explanations. Bottom: AI support with explanations.

The two experiments covered two different types of tasks (see Figure 3.1): the text-based profession classification task [P5] which we also used in [P4], as described in Section 3.1.1, and an image recognition task [P7]. Both experiments had one condition where the AI only gave a recommendation, and one condition where it in addition gave feature-based explanations that are typical for the respective task. In [P5], we conducted an online experiment with 200 participants. We measured decision difficulty through the agreement among participants, where lower agreement indicated higher decision difficulty. In [P7], we used electroencephalography (EEG) with 34 participants to measure mental workload as an indicator for decision difficulty. Both approaches had complementary strengths and weaknesses, as shown in Table 3.2.

We found in both studies that participants' reliance on AI was significantly less appropriate in more difficult decisions. In [P5], explanations helped to mitigate overreliance for easy decisions, but became less effective with increasing decision difficulty, with a tendency to even increase overreliance in the most difficult decisions. In [P7], the explanations had no significant effect, but we found indications that in principle, they could be helpful for easy decisions, but not for difficult ones.

We conclude that recommendation-centric support is helpful for easy decisions, where humans often know the answer themselves or at least have an intuition about the decision.

Table 3.2: Advantages and disadvantages of the two approaches used in this thesis for measuring decision difficulty.

	Advantages	**Disadvantages**
Agreement-based measure [P5]	Assesses decision difficulty on a continuous scale.	Does not account for subject dependency of decision difficulty.
EEG-based measure [P7]	Measures decision difficulty as subjectively experienced by each participant.	Only allows for binary distinction between easy and difficult decisions.

Here, recommendations and explanations can occasionally provide a useful second opinion. However, for difficult decisions where humans are highly uncertain about the correct answer, recommendation-centric support may be inappropriate, since the end-to-end recommendations and the technical explanations mostly do not help with forming a better intuition about the task. Consequently, decision makers have few other clues to work with than the recommendation and are likely to overrely on the AI.

Author contributions for [P5]: This publication is the result of the master thesis of my student Felicitas Buchner. I provided the study idea, conducted the statistical analysis in the form found in the publication, and wrote the paper. Felicitas designed the study under my guidance, implemented the machine learning model, the explanations, and study interface, and conducted the study. Andreas Butz contributed to the statistical analysis. Yuanting Liu and Andreas provided guidance and feedback and edited the final version of the submission.

Author contributions for [P7]: I contributed the study idea for this publication and led the collaboration. The experiment design was a joint effort, with Mariam Hassib and I focusing on the requirements from the perspective of AI-assisted decision-making, while the colleagues from TUBITAK—Seniha Ketenci Argın, Mustafa Baha Bilen, Doğan Urgun, and Sencer Melih Deniz—ensured that the experiment was compatible with the constraints of EEG data acquisition. I trained the neural network and implemented the explanations. The study interface and EEG data acquisition pipeline were implemented by the colleagues from TUBITAK, who also conducted the experiment and derived mental workload measures from the EEG data. Mariam and I performed the statistical analysis. Writing was a joint effort by Mariam, Doğan, Sencer, and me. Yuanting Liu provided guidance and feedback and edited the final submission.

3.1.3 Helpful When Correct, but a Burden When Wrong

This section is based on the following publications:

> *"**Pilot attitudes toward AI in the cockpit: implications for design**" [P1],*
>
> *"**Resilience through appropriation: Pilots' view on complex decision support**" [P3].*

To understand what hinders the adoption of AI-assisted decision-making in high-stakes domains, it is important to study real-world use cases in addition to experiments like those in Sections 3.1.1 and 3.1.2. We considered a use case from commercial aviation, which will be explained in more detail in Section 3.2.2. We conducted an initial exploratory interview study with four professional pilots [P1] to understand their perspective on the introduction of AI in the cockpit. We found that pilots' primary concern is not the black box nature of AI systems, as is often assumed when explanations are added to AI recommendations. Our participants were more concerned that AI may not be able to properly handle complex situations. A recurring example was that aviation relies on well-defined procedures that cover the majority of events; but in extraordinary situations, pilots may need to flexibly deviate from usual procedures. Our interviewees feared that an AI system may lack this flexibility and judgment, and become a burden rather than a help in such situations.

These findings highlight an important requirement for AI systems in high-stakes applications that is often not sufficiently addressed: Imperfect AI outputs, which are unavoidable in a complex domain such as aviation, should have minimal negative impact on users, a property Gu et al. [35] call *AI-resilient*. Recommendation-centric support tends to violate this requirement especially for complex decisions, as discussed by participants in a second study [P3], which will be explained in more detail in Section 3.2.2. While pilots stated that they would not blindly trust AI, they rejected the notion that they have to question case-by-case whether the AI recommendation is correct or not. Yet, this is the assumption driving most research on appropriate reliance in AI-assisted decision-making. Pilots emphasized that such a system would burden them more than it would help.

The seeming contradiction of not blindly trusting AI, but also not wanting to question it on a case-by-case basis, points to what makes recommendation-centric support difficult to work with: It diverts users' cognitive resources away from their own decision-making process and toward the review of a recommendation in which they have not been involved. Instead of augmenting users' decision-making process, reviewing the recommendation creates extra effort, especially when the decision is a complex one.

Author contributions for [P1]: The study idea was mine, and I also conducted the interviews, analyzed the data, and wrote the paper. Yuanting Liu and Heinrich Hußmann provided guidance and feedback and edited the final version of the submission.

Author contributions for [P3]: Given in Section 3.2.2, where the publication is described in more detail.

3.2 Exploration of Alternatives to Recommendation-Centric Support

> ***RQ2:*** *How can AI support high-stakes decisions without being recommendation-centric to overcome the limitations of recommendation-centric support?*

Similar to Section 3.1.3, studies of AI-assisted decision-making in real-world applications, such as in healthcare [9, 40, 98], social work [46], or sales [8], frequently find how recommendation-centric support does not fit well into users' decision-making process. However, concrete examples for alternative, *not* recommendation-centric ways to support human decision-making are rare; the few existing examples are mostly found in healthcare [54, 99, 105]. It is even less clear how such alternative designs compare to recommendation-centric support.

For RQ2, my objective was therefore to identify concrete ways for AI decision support that are not focused on recommendations, and compare them against typical recommendation-centric support. To this end, I studied a use case from commercial aviation.

3.2.1 Backward and Forward Reasoning

This section is based on the publication

> *"**Forward reasoning decision support: toward a more complete view of the human-AI interaction design space**" [P2].*

In this position paper, we laid out the conceptual groundwork for the following studies described in Sections 3.2.2 and 3.2.3. We argued that recommendation-centric support introduces a secondary task on top of users' primary decision-making task, namely the review of the AI's recommendation. This review task requires users to reason *backward* from the AI-recommended end result, which is error-prone and leads to inappropriate reliance [91]. In addition, for complex decisions, reasoning backward to validate an AI recommendation can be an effortful "research task", as Burgess et al. [9] put it. Users are unlikely to accept this extra effort on top of their demanding primary task [9, 40] [P3].

As an alternative, we proposed that rather than trying to solve the task for users end-to-end via a decision recommendation, AI could provide more incremental support. The goal is to avoid introducing a secondary task, and instead to support users to reason *forward* through their primary decision-making task.

On the surface, this incremental support seems less ambitious than recommendation-centric support because it appears to do less for users. Our key insight was that in reality, recommendation-centric support does not necessarily lead to less effort than more incremental support. Rather, the difference is that recommendation-centric support induces backward reasoning, while incremental support encourages forward reasoning, which can potentially

address the limitations of recommendation-centric support as identified in Section 3.1. This insight guided the studies in the following sections, where we aimed to identify ways to support pilots' reasoning in a forward direction (Section 3.2.2) and to compare this approach to recommendation-centric support (Section 3.2.3).

Author contributions: The idea for this position paper was mine, and I also wrote the paper. Yuanting Liu and Heinrich Hußmann provided guidance and feedback and edited the final version of the submission.

3.2.2 Continuous Support

This section is based on the publication

> *"**Resilience through appropriation: Pilots' view on complex decision support**" [P3].*

For the studies in this and the following section, we considered *diversions* in commercial aviation as a real-world use case for AI-assisted decision-making. Diversions are when a flight cannot reach its planned destination and has to divert to another airport, e.g. due to a medical emergency on board, a technical failure, or adverse weather conditions at the destination. Diversion decisions are in the responsibility of the pilots and are one of the primary use cases where pilots can imagine AI support [94].

To understand how AI can assist with diversion decisions, we designed two click-dummy prototypes and discussed them with eight professional pilots. The *Global Suggestions* prototype represents a typical recommendation-centric concept, where the system ranks the surrounding airports and recommends up to three of them. To make the recommendations transparent, the underlying decision criteria speaking for and against a particular airport are provided with a color coding in a table view. The idea behind the *Local Hints* prototype (Figure 3.2) on the other hand is to give hints about potential constraints at the surrounding airports, without explicitly recommending an airport. Crucially, the system is meant to be permanently displayed to pilots, even in normal flight when no diversion is imminent. This is supposed to improve pilots' situation awareness and to support their reasoning in a forward direction. During an emergency, pilots can enter an emergency mode to get hints that are tailored to the specific situation.

We found that diversion decisions involve much more than merely the point at which pilots choose the diversion airport. Before this point, during normal flight, pilots constantly aim to maintain situation awareness to be prepared for a possible diversion. They also ensure that there is always a valid plan B to avoid running out of options. After the decision point, pilots continue to check that their current plan is valid. Recommendation-centric support disregards this process nature of diversion decisions and only addresses the point where the actual decision is made. Our findings reveal the importance of continuously supporting the entire decision-making process, as exemplified by our Local Hints concept.

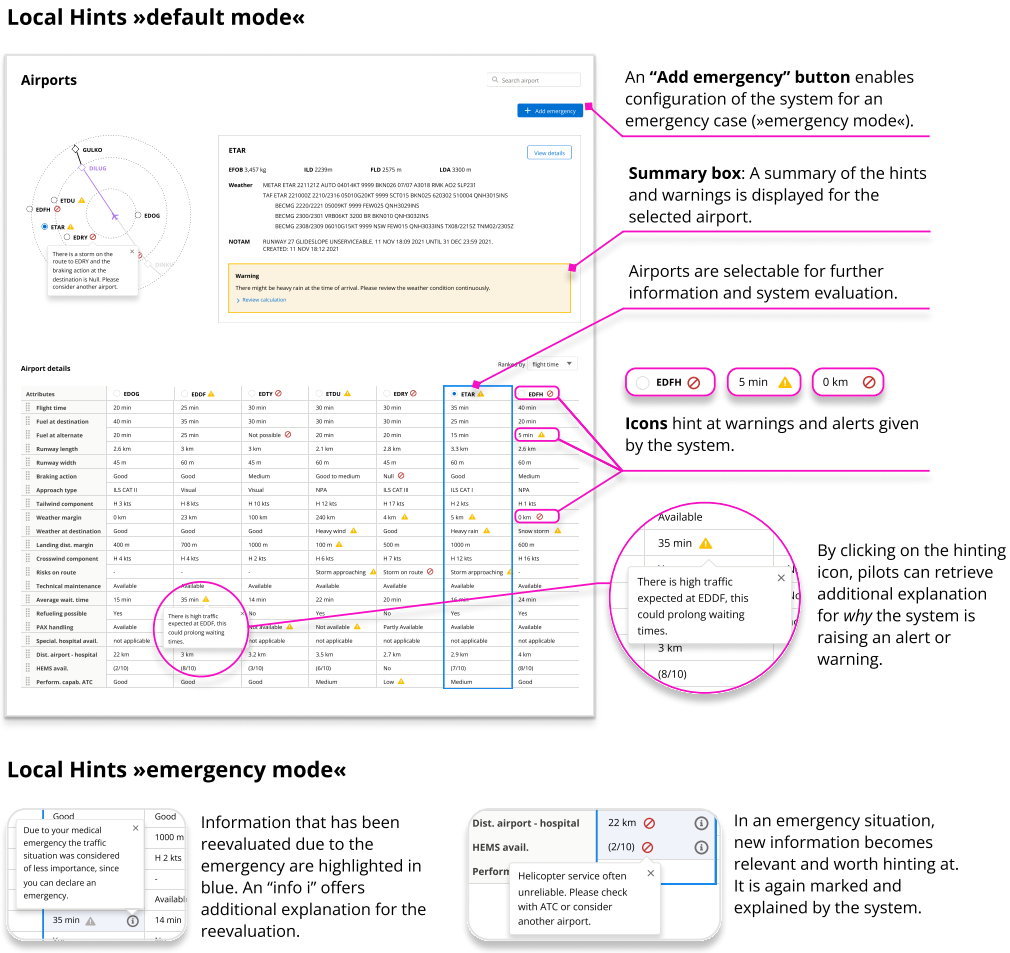


Figure 3.2: Overview of the Local Hints concept. Source: [P3].

We further found that pilots would not blindly trust an AI decision support tool, but they also do not want to be burdened with constantly questioning the correctness of the AI, as mentioned in Section 3.1.3. We found *appropriation* [23] to be a useful lens to make sense of this seeming contradiction. Pilots want to appropriate the AI support according to their current intention, such that the AI complements their reasoning, instead of creating a review task that is not integrated into pilots' decision-making. A more open form of support like the continuous provision of local hints might be easier to appropriate than closed, end-to-end recommendations.

Author contributions: Cara Storath and I contributed equally to this publication. Cara provided the study idea, and I contributed the writing of the paper. The study design and execution, prototype design, and data analysis were joint efforts by the two of us. Yuanting Liu and Andreas Butz provided guidance and feedback and edited the final version of the submission.

3.2.3 Comparison With Recommendation-Centric Support

This section is based on the publication

> ***"Beyond recommendations: From backward to forward AI support of pilots' decision-making process"*** *[P6].*

The results described in Section 3.2.2 showed that continuous support through local hints is a promising approach to support diversion decisions that is viewed favorably by pilots. However, it remained unclear how this approach compares to recommendation-centric support. Clarifying this question by conducting an empirical comparison was the aim of the work described in this section. We designed four versions of an AI-based diversion assistance tool for this mixed-methods study (see Figure 3.3): a recommendation-centric version (*Rec*), similar to the Global Suggestions concept described in Section 3.2.2; a continuous support version (*Cont*), similar to the Local Hints concept described in Section 3.2.2; a combined version that continuously provides local hints during normal flight and recommendations during emergency (*Rec+Cont*); and a baseline version that only presents surrounding airports and their data without AI elements.

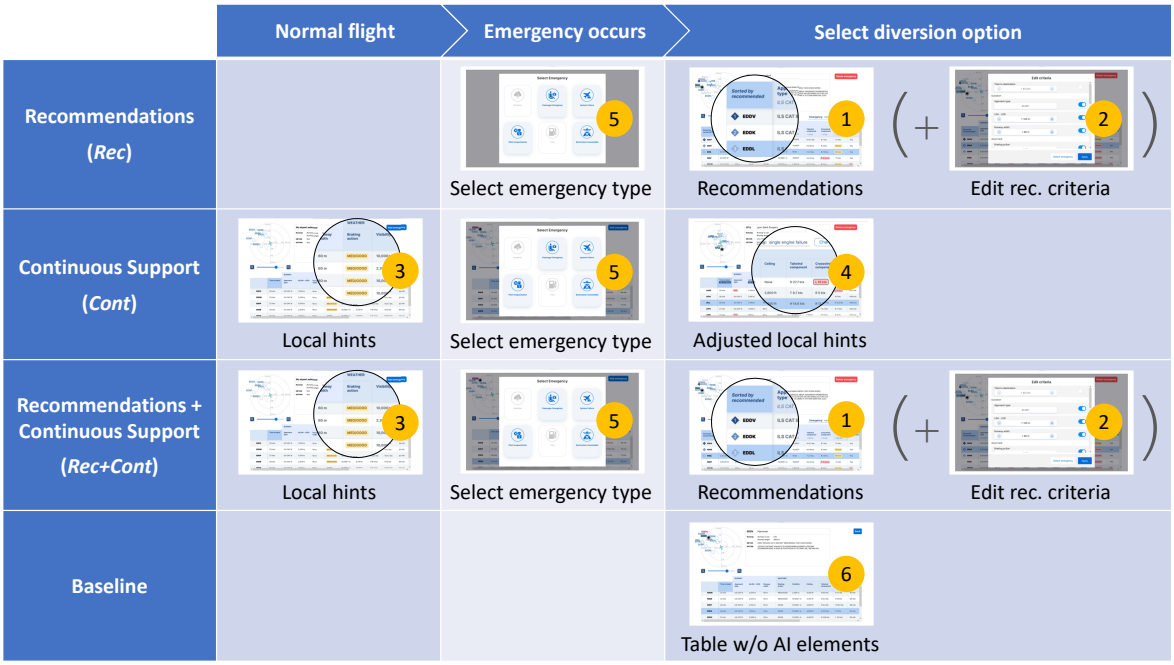


Figure 3.3: Study conditions for empirical comparison between recommendation-centric and continuous support. Source: [P6].

The study was based on three flight simulator scenarios that we validated and refined through a focus group with four professional pilots. Of particular interest was the third scenario, in which one airport—Hanover—would appear to be the best diversion option based on the AI evaluation. However, due to heavy traffic at that airport, which the AI does not consider in

its evaluation, Hanover would not be an optimal choice. We recruited 32 professional pilots for the study, with the diversion assistance version as between-subjects variable. Each pilot had to complete all three scenarios.

We found that the *Rec* version pushed pilots to reason backward, resulting in strong overreliance, as significantly more pilots chose Hanover in the third scenario than with the baseline version. In contrast, continuous support helped pilots to reason more in a forward direction, even when combined with recommendations. Consequently, pilots using the *Cont* and *Rec+Cont* versions where more likely than with the *Rec* version to think beyond the limits of the system in the third scenario and avoid Hanover. However, pilots were prone to fall back into backward reasoning in case of disruptions between normal flight and emergency, which also showed in that *Cont* and *Rec+Cont* participants were still more likely to choose Hanover than those using the baseline version.

Surprisingly, the *Rec* version did not lead to faster decisions than the *Cont* version, as the efficiency gain of directly receiving a recommendation was negated by the need to review the recommendation (backward reasoning). Yet, recommendations did lead to faster decisions in the second scenario with the *Rec+Cont* version, as the continuous support during normal flight allowed pilots to prepare a plan, which was then confirmed by the recommendation. As the recommendation fitted into pilots' reasoning (forward reasoning), there was no need to spend extra effort and time to review it.

Participants' statements in the exit interviews further revealed that pilots' biggest challenge during diversions is not to choose a suitable airport once they have the necessary information. Rather, the challenge is to gather and to integrate the information from multiple sources. Hence, even the baseline version was praised by participants since it conveniently displays the information at a glance. But due to the overwhelming amount of information, participants welcomed AI elements like recommendations or local hints to help them focus on what is relevant. However, recommendations were controversial among participants since some felt that recommendations would disengage them from the decision-making process. Continuous provision of local hints on the other hand was unanimously seen as positive.

Taken together, the results show that continuous support has advantages over the recommendation-centric support paradigm in terms of appropriate reliance, efficiency, and user acceptance. Recommendations can be beneficial, but rather than being the primary functionality of the AI system, recommendations should be embedded such that users engage with them while reasoning forward.

Author contributions: I provided the study idea and designed and conducted the main study. Sebastian Feger had the idea of the focus group and we conducted it together. Lucas Dullenkopf supported me with the design and implementation of the flight simulator scenarios. I designed and implemented most of the diversion assistance system, with support from my students Rulu Liao (interface design) and Lou Süsslin (implementation). I performed the data analysis, with Sebastian supporting as a second coder for the qualitative analysis. I also contributed the writing of the paper and the *process-oriented support* framework described

in Section 3.2.4. Sebastian, Yuanting Liu and Andreas Butz provided guidance and feedback and edited the final version of the submission.

3.2.4 Piecing It All Together: Process-Oriented Support

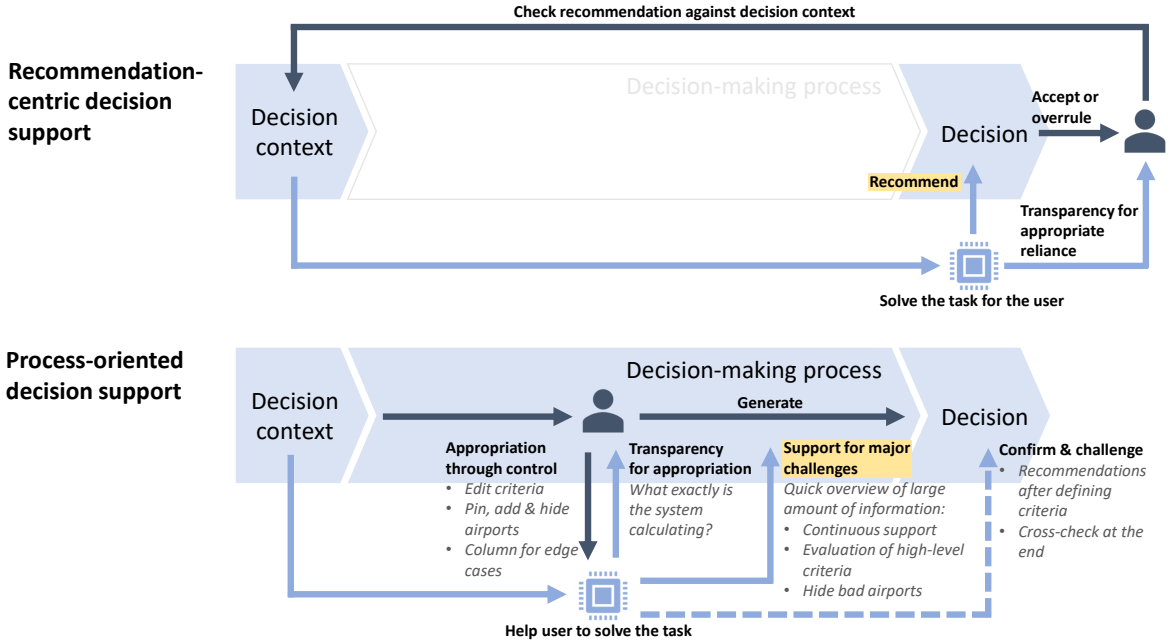


Figure 3.4: Comparison between recommendation-centric and process-oriented decision support. The yellow highlights signify the primary role of the AI in both paradigms. The dashed arrow indicates that recommendations are optional in process-oriented support. The italicised text describes examples for concrete implementations of process-oriented support for the diversion use case. Source: [P6].

Based on the results of Sections 3.2.1–3.2.3, we propose *process-oriented support* as a framework for designing AI decision support that is not recommendation-centric (see Figure 3.4). The framework is not limited to diversions or aviation, but meant to be widely applicable for AI-assisted decision-making in high-stakes applications, where it is crucial that humans remain fully in control. Continuous support can be seen as a concrete implementation of process-oriented support for the diversion use case. Figure 3.4 shows additional possibilities to implement process-oriented support, based on participants' suggestions in [P6].

Core to process-oriented support is to keep users engaged in the decision-making process in a forward direction. The primary functionality of the AI system is not to recommend a finished solution, but to provide support for the major challenges in users' decision-making process, which depend on the concrete application. In the diversion use case for instance, pilots' major challenge is to obtain a quick overview of a large amount of information. The goal is to help users make the decision, rather than making the decision for users.

To make the system resilient against imperfect AI inferences, a key component of the framework is to enable users to appropriate the system according to their momentary intentions. Appropriation can be facilitated through both transparency and control. Transparency is required to help users understand how well the system actions are aligned with their current intentions. As we discuss in more detail in [P6], this differs from the common notion of transparency for reliance calibration. Control is needed to steer the system when its actions do not fully align with users' intentions. This prevents a situation where the system is useful when it is correct, but useless when not. Note that this notion of control differs from interactive machine learning [3, 25], where users provide feedback that the system can learn from. Control for appropriation on the other hand is about users' momentary intentions that are heavily context-dependent and not necessarily learnable.

Process-oriented support might be able to address all of the limitations of recommendation-centric support identified in Section 3.1. In Section 3.1.1, we found that with recommendation-centric support, overreliance tends to increase over time, likely because users become less engaged as they familiarize with the AI system. This might be less of a problem with process-oriented support. Since the AI does not make the decision for users, users necessarily have to remain engaged in the decision-making process.

The results in Section 3.1.2 showed that users are able to rely appropriately on recommendation-centric support in easy decisions, but not in hard decisions where they are highly uncertain about the correct answer. The problem is that recommendation-centric support often does not help users to reduce their uncertainty and to develop a better intuition for the decision at hand. With process-oriented support, the decision-making is led by users' intuition and intentions. Ideally, any AI inference should build upon and help users develop their intuition about the decision task. In turn, this might help users to make better sense of how the AI's outputs fit into the decision context, leading to more appropriate reliance, as with the *Cont* and *Rec+Cont* variants in Section 3.2.3.

Lastly, in Section 3.1.3, we found that recommendation-centric support can be a burden for users due to the need to review the recommendations, which is especially effortful when the decision is complex. The problem is that users are not involved in the generation of end-to-end recommendations, meaning they have to reconstruct the recommendation from an external perspective. With process-oriented support, the aim is to integrate AI inferences into users' own decision-making process, which minimizes the extra effort of reviewing AI outputs. We observed this in Section 3.2.3 with the *Rec+Cont* version, where the recommendation served as a confirmation of users' own reasoning and led to a faster decision, while in the *Rec* version, pilots had to first review the recommendation, which slowed them down.

Chapter 4

Discussion

4.1 Related Concepts and Ideas

While the majority of research in AI-assisted decision-making takes a recommendation-centric approach, there is also a small, but growing body of work that explores alternative approaches. For instance, Lindvall et al. [54] designed a system for cancer assessment which, instead of marking images as containing cancer or not, helps pathologists to quickly identify and navigate to potentially interesting areas of an image to review. Zhang et al. [105] redesigned an existing tool for sepsis diagnosis to make actionable recommendations such as specific laboratory tests that can reduce the uncertainty of the diagnosis, instead of classifying whether a patient is septic or not. Studies like these give concrete examples for how AI can support human decision-making by other means than end-to-end recommendations and demonstrate the viability of these solutions. However, they represent point designs that can be hard to apply to other settings, and it is unclear how their effectiveness compares to recommendation-centric support.

In an attempt to move beyond point designs, Miller [60] proposed the concept of *Evaluative AI*, where the goal is to help users generate and evaluate different hypotheses, while leaving the choice between the candidate hypotheses to users. Evaluative AI can be considered a special case of process-oriented support, specifically for decision tasks where the main challenge is to compare several hypotheses, such as making medical diagnoses. So far, the concept has only been proposed as a position paper, without implemented systems and empirical studies to back it up.

Lately, another stream of work that aims to overcome the limitations of recommendation-centric support emerged under the term *human-centered explainable AI* (HCXAI) [27]. The goal with HCXAI is to shift away from explaining how the AI model technically generated its recommendation, toward providing explanations that are meaningful to the decision task itself. For example, Yang et al. [100] supplemented AI recommendations with references to evidence in the medical literature that supports or contradicts the recommendation. Ehsan

et al. [26] explored including socio-organizational factors into explanations in a sales use case. These approaches give users more context for their decision and help them to better reconcile the AI recommendation with the decision context. The recommendations become one part of a more holistic support instead of the primary functionality of the system. I consider HCXAI and approaches like process-oriented support that aim to identify alternative roles for AI support to be complementary on the path toward AI decision support that truly complements and augments human reasoning.

4.2 Positioning Within the Broader Human-Centered AI Landscape

Beyond AI-assisted decision-making, this thesis can be situated within the more general research field of *human-centered AI* (HCAI) [81]. Calls for developing AI that is human-centered have become commonplace [48, 52, 63, 76, 95, 96], but the understanding of what human-centeredness means in AI systems can diverge considerably [16]. For some, "human-centered" may primarily mean interacting with humans and seeking technical solutions to address human needs, while others emphasize the importance of human-centered design of AI applications. With the work in this thesis, I strongly subscribe to the latter understanding of HCAI, mostly informed by my study of a complex, high-stakes domain like aviation, where tolerance for mistakes is low, but potential for mistakes is high. In such a domain, addressing shortcomings of AI systems with more AI inferences (e.g., detection of user states, or interpretation of user context) is likely to introduce even more sources of error, especially given that today's machine learning-based AI systems are essentially pattern recognizers without real understanding of our world [30].

It appears more promising to me, at least in the short term, but likely also in the long term [93], to address the imperfections of AI through careful design informed by a deep understanding of human needs. The high-level ideas underlying the work in this thesis have been well articulated by Shneiderman [81]. First, when designing AI applications, it is helpful to break away from the urge to replicate human capabilities with AI. In successful applications, AI frequently assumes a well-defined, clearly scoped functionality that can differ significantly from human work [80]. Second, the widespread call for human-in-the-loop solutions reflects a thinking where AI systems are designed with as much autonomy as possible as a starting point; the human is then placed back into the loop to act as a fallback for the AI's limitations [29]. This often results in AI-centered systems, despite claims of pursuing human-centered AI, as Woods eloquently put it: "*The road to technology-centered systems is paved with user-centered intentions.*" [73]. As Shneiderman argues, it may be more useful to think of *AI-in-the-loop*, where the starting point is the human work and how AI can fit in there [81]. Both ideas—AI functionality beyond replicating human capabilities, and AI-in-the-loop—are reflected in the continuous support concept in Sections 3.2.2 and 3.2.3. Rather than mimicking a human co-pilot, the system continuously displays helpful information in a way that fits into pilots' decision-making process.

4.3 On Trust and Reliance

The attentive reader might have noticed that I did not address trust in this thesis; instead, I focused on reliance. As trust is one of the most prominent constructs in AI-assisted decision-making [90] and HCAI in general [20, 82], a dedicated section about my reasons for this decision seems appropriate. The first reason, as argued by Dorsch and Deroy [24], is that in high-stakes applications, trustworthiness may be a misguided objective for the development of AI systems, since AI does not posses moral agency and accountability. Instead, reliability is what is required. The second reason is that a central aspect of common definitions of trust is a situation of vulnerability as precondition for trust to exist [90]. Yet, I argue that for high-stakes applications, one should strive to minimize users' vulnerability to imperfections of AI, which implies that essentially trust should not be required.

To elaborate on the second reason, I suggest that rather than asking users to trust AI, high-stakes AI applications should be resilient against AI mistakes (see Section 3.1.3). Ideally, the cost to the user should be minimal when AI makes mistakes. If that is not possible, users should be able to judge when they want to rely on AI to perform certain tasks, based on an understanding of whether the cost of imperfect AI performance is acceptable in a specific situation. For instance, if an imperfect result is good enough or easily correctable for the user, the user may still want to rely and iterate on the imperfect AI output. Conversely, if the cost of an AI mistake is too high, the user should be able to easily recognize this and dismiss the AI. This puts users in control and never leaves them vulnerable to the possibility of AI mistakes. To put it a little provocatively: trust is good, control is better.

4.4 Limitations and Future Work

When exploring alternatives to recommendation-centric support (Section 3.2), I mostly focused on the *roles* AI could assume in AI-assisted decision-making. While the diversion assistance system provided transparency and control mechanisms, they were not the focus of my work. Yet, the results emphasized the importance of enabling appropriation of AI tools through transparency and control. Future work should investigate further how to facilitate appropriation in AI-assisted decision-making.

Another obvious direction for future work is to study how process-oriented support applies to other use cases and domains than diversions in aviation. Related studies, especially in the healthcare domain, point to very similar concerns to those expressed by pilots around recommendation-centric support [9, 40], as well as users' desire for AI to support more intermediary stages of their decision-making process [101, 105]. Therefore, and given that the process-oriented support framework is kept fairly general, I argue that the framework also applies to other high-stakes decision tasks. While I am confident that the framework can already be useful in its general form presented in Section 3.2.4 by emphasizing appropriation and supporting the decision-making process, future work can explore how it can be con-

cretized for specific use cases. In addition, there is a wealth of other helpful AI support roles that future work can explore, as other decision tasks may offer different opportunities for AI support. Future work should further aim to better understand when either recommendation-centric or process-oriented support is preferable. For example, the results in Section 3.1.2 suggest that for easy decisions, recommendation-centric support may very well be an effective form of support.

As a third line of future work, it is important to recognize that in this thesis, I take a highly positive view of human expertise. However, as evidenced by the rich body of work on human cognitive heuristics and biases [43, 87], human decision-making can be flawed, which can also manifest when humans take the lead in interactions with AI [62, 70]. An interesting research question, therefore, is how AI can help mitigate human decision errors. For instance, AI can be used to detect and understand sources of human error [104]. Research on human decision-making has also uncovered conditions delineating when humans are likely to make mistakes or to display true expertise [44]. Building on these results is crucial to understand how AI can mitigate human errors while promoting human expertise in AI-assisted decision-making.

Finally, the work in this thesis remains relevant with the current surge in large language models and generative AI. On the one hand, the recent breakthroughs in generative AI could enable powerful novel capabilities for AI-assisted decision-making [28]. On the other hand, generative AI could aggravate the problems identified in this thesis, for example through end-to-end recommendations for even more complex problems [83], or through even more complex failure modes [86]. It remains to be seen how the potential of generative AI can be effectively leveraged for AI-assisted decision-making, but I expect that appropriation and aiming to help users solve their task, rather than solving the task for them, will remain important strategies.

Chapter 5

Conclusion

In this thesis, I investigated the limitations of recommendation-centric support in minimizing the impact of imperfect AI performance in AI-assisted decision-making. I further explored how reconsidering the role of AI from solving the task for users, to helping users to solve the task, is a promising path to benefit from imperfect AI in high-stakes decisions. Rather than sophisticated algorithms, this approach, which I captured in my framework for process-oriented support, first and foremost requires a deep understanding of users' decision-making processes and challenges.

My research contributes to the growing, but still small, body of work that challenges the assumption that AI-assisted decision-making should be recommendation-centric. In contrast to related work, which is mostly concerned with the healthcare domain, I studied a use case from aviation, a domain that is rarely studied in HCI. By finding that pilots share many of the same concerns as clinicians, I contribute to the generalizability of recent results beyond the healthcare domain. Related work is also often formative, in that studies uncover problems with recommendation-centric support and identify opportunities for alternative forms of support at an abstract level, but concrete solutions that exploit these opportunities are rare. I contribute such a concrete solution for the diversion use case in the form of continuous support, and captured the generalizable elements of it in the process-oriented support framework. Most importantly, I conducted an empirical comparison between continuous and recommendation-centric support, contributing some of the first evidence that alternative forms of support may be more effective at supporting human decision-making than recommendation-centric support in certain use cases.

I argue that my results are also relevant for the broader field of human-centered AI. Despite its impressive capabilities, AI is always imperfect. We cannot ignore this and design AI-driven systems as if AI was perfect. Yet, this seemingly trivial statement is often not given enough attention, which usually shows in that an AI system is designed to solve users' tasks for them. Solving tasks for users is desirable if you can always solve them perfectly, but that is rarely the case. Due to the focus on the expected gains when AI performs well, the potential for imperfect AI performance becomes an afterthought, addressed by assigning

users the difficult role of a fallback for AI. In AI-assisted decision-making, this approach is embodied by the predominant recommendation-centric support paradigm, and the goal of calibrating users' reliance on end-to-end decision recommendations. A similar tendency can be observed for other human-centered AI applications. Based on my research, I suspect that for other AI applications as well, it might be helpful to explore more carefully designed incremental AI support that keeps users engaged in their tasks, rather than using AI to solve tasks end-to-end by default.

REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, August 2020. DOI: 10.1109/MC.2020.2996587.

[2] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, November 2020. DOI: 10.1186/s12911-020-01332-6.

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: the role of humans in interactive machine learning. *AI Magazine*, 35(4): 105–120, December 2014. DOI: 10.1609/aimag.v35i4.2513.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, July 2016. DOI: 10.48550/arXiv.1606.06565. arXiv:1606.06565v2 [cs.AI].

[5] Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, 159:107197:1–107197:15, July 2023. DOI: 10.1016/j.infsof.2023.107197.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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, pp. 81:1–81:16, Yokohama, Japan, May 2021. ACM. DOI: 10.1145/3411764.3445717.

[7] Astrid Bertrand, James R. Eagan, and Winston Maxwell. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and*

Transparency, FAccT '23, pp. 943–958, Chicago, IL, USA, June 2023. ACM. DOI: 10.1145/3593013.3594053.

[8] Jeanette Blomberg, Aly Megahed, and Ray Strong. Acting on analytics: accuracy, precision, interpretation, and performativity. *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1):281–300, 2018. DOI: 10.1111/1559-8918.2018.01208.

[9] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. Healthcare AI treatment decision support: design principles to enhance clinician adoption and trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 15:1–15:19, Hamburg, Germany, April 2023. ACM. DOI: 10.1145/3544548.3581251.

[10] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics*, ICHI 2015, pp. 160–169, Dallas, TX, USA, October 2015. IEEE. DOI: 10.1109/ICHI.2015.26.

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21, April 2021. DOI: 10.1145/3449287.

[12] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Susan A. Murphy, and Krzysztof Z. Gajos. Towards optimizing human-centric objectives in AI-assisted decision-making with offline reinforcement learning, April 2024. arXiv:2403.05911v2 [cs.HC].

[13] Federico Cabitza, Andrea Campagner, and Carla Simone. The need to move away from agential-AI: empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies*, 155:102696:1–102696:11, November 2021. DOI: 10.1016/j.ijhcs.2021.102696.

[14] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):136:1–136:21, April 2023. DOI: 10.1145/3579612.

[15] Longbing Cao. AI in finance: challenges, techniques, and opportunities. *ACM Computing Surveys*, 55(3):64:1–64:38, February 2022. DOI: 10.1145/3502289.

[16] Tara Capel and Margot Brereton. What is human-centered about human-centered AI? A map of the research landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 359:1–359:23, Hamburg, Germany, April 2023. ACM. DOI: 10.1145/3544548.3580959.

[17] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, October 2019. DOI: 10.1177/0022243719851788.

[18] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 370:1–370:32, October 2023. DOI: 10.1145/3610219.

[19] Lingwei Cheng and Alexandra Chouldechova. Overcoming algorithm aversion: a comparison between process and outcome control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 756:1–756:27, Hamburg, Germany, April 2023. ACM. DOI: 10.1145/3544548.3581253.

[20] Erin K. Chiou and John D. Lee. Trusting automation: Designing for responsivity and resilience. *Human Factors*, 65(1):137–165, February 2023. DOI: 10.1177/00187208211009995. Publisher: SAGE Publications Inc.

[21] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 509:1–509:12, Honolulu, HI, USA, April 2020. ACM. DOI: 10.1145/3313831.3376638.

[22] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. DOI: 10.1037/xge0000033.

[23] Alan Dix. Designing for appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers*, volume 2 of *BCS-HCI '07*, pp. 27–30, Lancaster, UK, September 2007. BCS Learning & Development Ltd. DOI: 10.14236/ewic/HCI2007.53.

[24] John Dorsch and Ophelia Deroy. "Quasi-metacognitive machines: why we don't need morally trustworthy AI and communicating reliability is enough". *Philosophy & Technology*, 37(2):62, May 2024. DOI: 10.1007/s13347-024-00752-w.

[25] John J. Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 8(2): 8:1–8:37, July 2018. DOI: 10.1145/3185517.

[26] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 82:1–82:19, Yokohama, Japan, May 2021. ACM. DOI: 10.1145/3411764.3445188.

[27] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, pp. 109:1–109:7, New Orleans, LA, USA, April 2022. ACM. DOI: 10.1145/3491101.3503727.

[28] Eva Eigner and Thorsten Händler. Determinants of LLM-assisted decision-making, February 2024. arXiv:2402.17385 [cs].

[29] Madeleine Clare Elish. Moral crumple zones: cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, 5:40–60, March 2019. DOI: 10.17351/ests2019.260.

[30] Mica R. Endsley. Ironies of artificial intelligence. *Ergonomics*, 66(11):1656–1668, November 2023. DOI: 10.1080/00140139.2023.2243404. Publisher: Taylor & Francis.

[31] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1362–1374, Seoul, Republic of Korea, June 2022. ACM. DOI: 10.1145/3531146.3533193.

[32] Raymond Fok and Daniel S. Weld. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 45(3):317–332, July 2024. DOI: 10.1002/aaai.12182. Publisher: John Wiley & Sons Ltd.

[33] Krzysztof Z. Gajos and Lena Mamykina. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces*, IUI '22, pp. 794–806, Helsinki, Finland, March 2022. ACM. DOI: 10.1145/3490099.3511138.

[34] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 50:1–50:24, November 2019. DOI: 10.1145/3359152.

[35] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. An AI-resilient text rendering technique for reading and skimming documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 898:1–898:22, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3642699.

[36] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, August 2018. DOI: 10.1145/3236009.

[37] Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 221–236, Rio de Janeiro, Brazil, June 2024. ACM. DOI: 10.1145/3630106.3658901.

[38] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):276:1–276:29, October 2023. DOI: 10.1145/3610067.

[39] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30(5):71:1–71:29, September 2023. DOI: 10.1145/3534561.

[40] Maia Jacobs, Jeffrey He, Melanie F Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof 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*, CHI '21, pp. 659:1–659:14, Yokohama, Japan, May 2021. ACM. DOI: 10.1145/3411764.3445385.

[41] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1):108:1–108:9, June 2021. DOI: 10.1038/s41398-021-01224-x.

[42] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, July 2018. DOI: 10.1016/j.bushor.2018.03.007.

[43] Daniel Kahneman. *Thinking, fast and slow*. Penguin Books, London, 2012.

[44] Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, 2009. DOI: 10.1037/a0016755. Publisher: American Psychological Association.

[45] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 52:1–52:18, New Orleans, LA, USA, April 2022. ACM. DOI: 10.1145/3491102.3517439.

[46] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. “Why do I care what’s similar?” Probing challenges in AI-assisted child welfare decision-making through worker-AI interface design concepts. In *Designing Interactive Systems Conference*, DIS ’22, pp. 454–470, Virtual Event, Australia, June 2022. ACM. DOI: 10.1145/3532106.3533556.

[47] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I’m not sure, but...": Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, Rio de Janeiro, Brazil, June 2024. ACM. DOI: 10.1145/3630106.3658941.

[48] Gary Klein, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, November 2004. DOI: 10.1109/MIS.2004.74.

[49] Vivian Lai and Chenhao Tan. 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, pp. 29–38, Atlanta, GA, USA, January 2019. ACM. DOI: 10.1145/3287560.3287590.

[50] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 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, pp. 1369–1385, Chicago, IL, USA, June 2023. ACM. DOI: 10.1145/3593013.3594087.

[51] John D. Lee and Katrina A. See. Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46 (1):50–80, 2004. DOI: 10.1518/hfes.46.1.50_30392.

[52] Q. Vera Liao and Jennifer Wortman Vaughan. AI transparency in the age of LLMs: a human-centered research roadmap. *Harvard Data Science Review*, (Special Issue 5): 1–40, May 2024. DOI: 10.1162/99608f92.8036d03b. Publisher: The MIT Press.

[53] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: a review of machine learning interpretability methods. *Entropy*, 23(1):18:1–18:45, December 2020. DOI: 10.3390/e23010018. Publisher: Multidisciplinary Digital Publishing Institute.

[54] Martin Lindvall, Claes Lundström, and Jonas Löwgren. Rapid assisted visual search: supporting digital pathologists with imperfect AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI ’21, pp. 504–513, College Station, TX, USA, April 2021. ACM. DOI: 10.1145/3397481.3450681.

[55] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):408:1–408:45, October 2021. DOI: 10.1145/3479552.

[56] Zhuoran Lu, Dakuo Wang, and Ming Yin. Does more advice help? The effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):217:1–217:31, April 2024. DOI: 10.1145/3653708.

[57] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 759:1–759:19, Hamburg, Germany, April 2023. ACM. DOI: 10.1145/3544548.3581058.

[58] John M. McGuirl and Nadine B. Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(4):656–665, December 2006. DOI: 10.1518/001872006779166334.

[59] Bryce McLaughlin and Jann Spiess. Designing algorithmic recommendations to achieve human-AI complementarity, May 2024. arXiv:2405.01484v1 [cs.HC].

[60] Tim Miller. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 333–342, Chicago, IL, USA, June 2023. ACM. DOI: 10.1145/3593013.3594001.

[61] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, pp. 340–350, College Station, TX, USA, April 2021. ACM. DOI: 10.1145/3397481.3450639.

[62] Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. The human in the infinite loop: a case study on revealing and explaining human-AI interaction loop failures. In *Proceedings of Mensch und Computer 2022*, MuC '22, pp. 158–168, Darmstadt, Germany, September 2022. ACM. DOI: 10.1145/3543758.3543761.

[63] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3):

391–437, February 2023. DOI: 10.1080/10447318.2022.2153320. Publisher: Taylor & Francis.

[64] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4):35:1–35:33, March 2022. DOI: 10.1145/3495013.

[65] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):102:1–102:15, November 2019. DOI: 10.1145/3359204.

[66] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-AI teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12089–12097, Vancouver, BC, Canada, June 2022. DOI: 10.1609/aaai.v36i11.21468. Number: 11.

[67] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 237:1–237:52, Yokohama, Japan, May 2021. ACM. DOI: 10.1145/3411764.3445315.

[68] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pp. 379–396, Sydney, NSW, Australia, March 2023. ACM. DOI: 10.1145/3581641.3584033.

[69] Andrew Prahl and Lyn Van Swol. Understanding algorithm aversion: when is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702, 2017. DOI: 10.1002/for.2464.

[70] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: the role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):83:1–83:22, April 2022. DOI: 10.1145/3512930.

[71] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 535:1–535:14, New Orleans, LA, USA, April 2022. ACM. DOI: 10.1145/3491102.3501967.

[72] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 25:1–25:17, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3641905.

[73] Nadine B. Sarter, David D. Woods, and Charles E. Billings. Automation surprises. In Gavriel Salvendy, editor, *Handbook of Human Factors & Ergonomics*, pp. 1926–1943. John Wiley & Sons, 2 edition, 1997.

[74] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 617–626, Oxford, United Kingdom, July 2022. ACM. DOI: 10.1145/3514094.3534128.

[75] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pp. 410–422, Sydney, NSW, Australia, March 2023. ACM. DOI: 10.1145/3581641.3584066.

[76] Albrecht Schmidt. Interactive human centered artificial intelligence: a definition and research challenges. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*, AVI '20, pp. 3:1–3:4, Salerno, Italy, October 2020. ACM. DOI: 10.1145/3399715.3400873.

[77] Philipp Schmidt and Felix Biessmann. Calibrating human-AI collaboration: impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction*, CD-MAKE 2020, pp. 431–449, Dublin, Ireland, August 2020. Springer International Publishing. DOI: 10.1007/978-3-030-57321-8_24.

[78] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 836:1–836:18, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3642621.

[79] Silvana Secinaro, Davide Calandra, Aurelio Secinaro, Vivek Muthurangu, and Paolo Biancone. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, 21(1):125, April 2021. DOI: 10.1186/s12911-021-01488-9.

[80] Ben Shneiderman. Design lessons from AI's two grand goals: human emulation and useful applications. *IEEE Transactions on Technology and Society*, 1(2):73–82, June 2020. DOI: 10.1109/TTS.2020.2992669.

[81] Ben Shneiderman. Human-centered artificial intelligence: three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124, 2020. DOI: 10.17705/1thci.00131.

[82] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, April 2020. DOI: 10.1080/10447318.2020.1741118. Publisher: Taylor & Francis.

[83] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interactions, February 2024. DOI: 10.48550/arXiv.2402.11364. arXiv:2402.11364 [cs].

[84] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong-ah (Kate) Jeong, Shin'ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris, and Jia Zhou. Seven HCI Grand Challenges. *International Journal of Human–Computer Interaction*, 35(14):1229–1269, August 2019. DOI: 10.1080/10447318.2019.1619259.

[85] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42, August 2019. DOI: 10.1177/0008125619867910.

[86] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 680:1–680:24, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3642902.

[87] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, September 1974. DOI: 10.1126/science.185.4157.1124. Publisher: American Association for the Advancement of Science.

[88] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):129:1–129:38, April 2023. DOI: 10.1145/3579605.

[89] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952:1–103952:24, September 2023. DOI: 10.1016/j.artint.2023.103952.

[90] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2):327:1–327:39, October 2021. DOI: 10.1145/3476068.

[91] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 601:1–601:15, Glasgow, Scotland, UK, May 2019. ACM. DOI: 10.1145/3290605.3300831.

[92] Xinru Wang and Ming Yin. 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, pp. 318–328, College Station, TX, USA, April 2021. ACM. DOI: 10.1145/3397481.3450650.

[93] David D. Woods. The risks of autonomy: Doyle's Catch. *Journal of Cognitive Engineering and Decision Making*, 10(2):131–133, June 2016. DOI: 10.1177/1555343416653562.

[94] Jakob Würfel, Boris Djartov, Anne Papenfuß, and Matthias Wies. Intelligent Pilot Advisory System: The journey from ideation to an early system design of an AI-based decision support system for airline flight decks. In *Human Factors in Transportation*, volume 95 of *AHFE 2023*, pp. 589–597, San Francisco, CA, USA, 2023. AHFE Open Acces. DOI: 10.54941/ahfe1003844.

[95] Wei Xu. Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, June 2019. DOI: 10.1145/3328485.

[96] Wei Xu and Zaifeng Gao. Applying HCAI in developing effective human-AI teaming: a perspective from human-AI joint cognitive systems. *Interactions*, 31(1):32–37, January 2024. DOI: 10.1145/3635116.

[97] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pp. 189–201, Cagliari, Italy, March 2020. ACM. DOI: 10.1145/3377325.3377480.

[98] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 4477–4488, San Jose, CA, USA, May 2016. ACM. DOI: 10.1145/2858036.2858373.

[99] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable AI: fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 238:1–238:11, Glasgow, Scotland, UK, May 2019. ACM. DOI: 10.1145/3290605.3300468.

[100] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 14:1–14:14, Hamburg, Germany, April 2023. ACM. DOI: 10.1145/3544548.3581393.

[101] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. Multimodal healthcare AI: Identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 444:1–444:22, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3642013.

[102] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–12, Glasgow, Scotland, UK, May 2019. ACM. DOI: 10.1145/3290605.3300509.

[103] Kun-Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10):719–731, October 2018. DOI: 10.1038/s41551-018-0305-z.

[104] J.D. Zamfirescu-Pereira, Jerry Chen, Emily Wen, Allison Koenecke, Nikhil Garg, and Emma Pierson. Trucks don't mean Trump: Diagnosing human error in image analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 799–813, Seoul, Republic of Korea, June 2022. ACM. DOI: 10.1145/3531146.3533145.

[105] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 445:1–445:18, Honolulu, HI, USA, May 2024. ACM. DOI: 10.1145/3613904.3642343.

[106] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 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, pp. 295–305, Barcelona, Spain, January 2020. ACM. DOI: 10.1145/3351095.3372852.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 26.09.2024

Zelun Tony Zhang