Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process

ZELUN TONY ZHANG, fortiss GmbH, Research Institute of the Free State of Bavaria, Germany and LMU Munich, Germany
SEBASTIAN S. FEGER, LMU Munich, Germany and TH Rosenheim, Germany
LUCAS DULLENKOPF, Airbus Defence and Space GmbH, Germany
RULU LIAO*, LMU Munich, Germany
LOU SÜSSLIN*, TU Wien, Austria
YUANTING LIU, fortiss GmbH, Research Institute of the Free State of Bavaria, Germany
ANDREAS BUTZ, LMU Munich, Germany

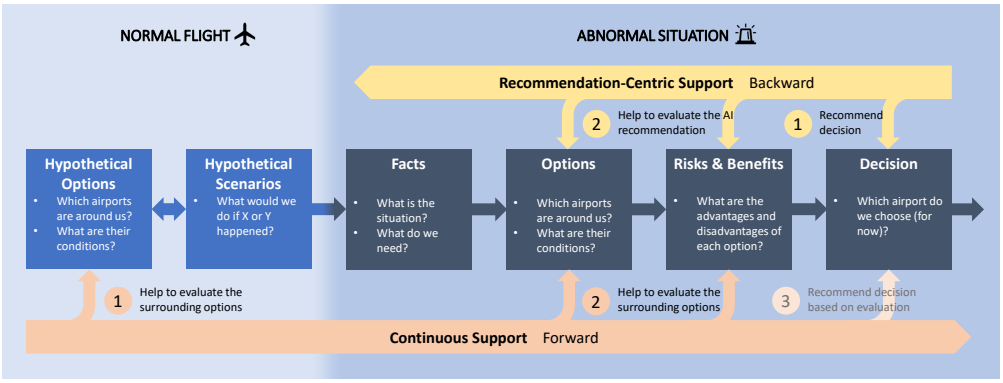


Fig. 1. Conceptual overview of the two decision support paradigms that we compare: *recommendation-centric support* and *continuous support*. While the former pushes pilots to reason backward from a decision recommendation, the aim of the latter is to support pilots in a forward direction. Instead of only popping up during an emergency, the system continuously helps pilots to evaluate their surroundings even during normal flight. The system does not give recommendations to avoid biasing pilots. But given continuous support, it may be possible to add recommendations in an emergency while still allowing pilots to reason forward, since pilots are already engaged with the system when they see the recommendations.

*The fourth and fifth authors contributed to this work while working at fortiss GmbH.

Authors' Contact Information: Zelun Tony Zhang, zhang@fortiss.org, fortiss GmbH, Research Institute of the Free State of Bavaria, Munich, Germany and LMU Munich, Munich, Germany; Sebastian S. Feger, LMU Munich, Munich, Germany and TH Rosenheim, Rosenheim, Germany, sebastian.feger@ifi.lmu.de; Lucas Dullenkopf, lucas.dullenkopf@airbus.com, Airbus Defence and Space GmbH, Manching, Germany; Rulu Liao, rulu.liao@campus.lmu.de, LMU Munich, Munich, Germany; Lou Süsslin, suesslin@pm.me, TU Wien, Vienna, Austria; Yuanting Liu, liu@fortiss.org, fortiss GmbH, Research Institute of the Free State of Bavaria, Munich, Germany; Andreas Butz, butz@ifi.lmu.de, LMU Munich, Munich, Germany.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2573-0142/2024/11-ART485
https://doi.org/10.1145/3687024

AI is anticipated to enhance human decision-making in high-stakes domains like aviation, but adoption is often hindered by challenges such as inappropriate reliance and poor alignment with users' decision-making. Recent research suggests that a core underlying issue is the recommendation-centric design of many AI systems, i.e., they give end-to-end recommendations and ignore the rest of the decision-making process. Alternative support paradigms are rare, and it remains unclear how the few that do exist compare to recommendation-centric support. In this work, we aimed to empirically compare recommendation-centric support to an alternative paradigm, continuous support, in the context of diversions in aviation. We conducted a mixed-methods study with 32 professional pilots in a realistic setting. To ensure the quality of our study scenarios, we conducted a focus group with four additional pilots prior to the study. We found that continuous support can support pilots' decision-making in a forward direction, allowing them to think more beyond the limits of the system and make faster decisions when combined with recommendations, though the forward support can be disrupted. Participants' statements further suggest a shift in design goal away from providing recommendations, to supporting quick information gathering. Our results show ways to design more helpful and effective AI decision support that goes beyond end-to-end recommendations.

CCS Concepts: • **Information systems** → **Decision support systems**; • **Human-centered computing** → **Interaction paradigms**.

Additional Key Words and Phrases: human-AI interaction, decision support tools, intelligent decision support, AI-assisted decision-making, human-AI decision-making, imperfect AI, decision support paradigms, continuous support, process-oriented support, aviation

ACM Reference Format:
Zelun Tony Zhang, Sebastian S. Feger, Lucas Dullenkopf, Rulu Liao, Lou Süsslin, Yuanting Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 485 (November 2024), 32 pages. https://doi.org/10.1145/3687024

1 Introduction

AI is projected to improve human decision-making in various high-stakes domains, such as healthcare [72], finance [11], or law enforcement [50]. Another domain is aviation [17], where AI is expected to not only increase decision efficiency, but also safety, as faulty decision-making is one of the main reasons for accidents in aviation [57], motivating research at the intersection between novel technologies and human factors [20]. One type of decisions in aviation are *diversion* decisions. A diversion is when a flight is unable to reach its planned destination, e.g. due to a technical failure, a medical emergency, or adverse weather conditions. It is the pilots' responsibility to decide on an alternate airport to divert to. While diversions are rare, they are very disruptive and costly for operations [40]. Poor diversion decisions can further increase the cost or even impact flight safety. Today, pilots have various tools to support them during diversions, but these are often cumbersome to use and not integrated with each other. Diversions are therefore one primary use case where pilots seek better support and where they can imagine AI assistance [68].

Yet, in spite of good machine performance, real-world adoption of AI is often difficult [55, 63]. In controlled studies, failure to achieve *complementary performance* is often observed [2, 22, 43], i.e., the combination of human and AI performs worse than one of them alone. This is the result of *inappropriate reliance*, including both *overreliance* (human relies on AI even when it is disadvantageous to do so) [2, 6, 8, 24, 32, 39, 43, 47, 56, 66] and *underreliance* (human rejects AI even when it would be beneficial to rely on it) [12–14, 49].

In practice, AI support often turns out less useful to decision makers than imagined [4, 34, 71]. A growing number of formative studies on real-world tasks with domain experts [1, 5, 31, 33–35, 59, 71, 73] try to understand what hinders effective use of AI decision support. A frequent problem is that AI decision support is usually designed to be *recommendation-centric*, where the

primary functionality of the system is to give end-to-end decision recommendations, i.e., the system suggests a possible end result straight from its input data. By directly jumping to the end result, these systems only support the very end of the decision-making [7], ignoring the entire process leading up to the decision [71, 73, 76].

While the limitations of recommendation-centric support are becoming increasingly more apparent, there is a notable lack of effective solutions to the identified problems. Current research is mostly limited to proposing alternatives for recommendation-centric support on a conceptual level [7, 9, 36, 45, 75], but concrete examples are few and far between, with most examples stemming from healthcare [10, 25, 42, 62, 70, 73]. Even fewer works evaluate these alternative support paradigms in comparison to recommendation-centric decision support.

One of the few studies exploring alternative decision support roles for AI is Zhang et al.'s work on diversion assistance [76]. Their key insight was that diversion decisions are not a *point*, but a *process* in which pilots take proactive actions. Even during normal flight, pilots constantly establish situation awareness (SA) and prepare a valid plan B, should an emergency occur. Continuously supporting these proactive actions via unobtrusive hints was found to be a promising support role for AI, but it remains unclear how effective it would be in practice, especially compared to the established recommendation-centric paradigm.

In this paper, we sought to empirically compare continuous against recommendation-centric support (see Figure 1) in terms of their effects on pilots' decision-making process, decision outcomes, and decision time in a realistic task setting. We conducted a mixed-methods study with 32 professional pilots, where pilots made a series of diversion decisions with either recommendations, continuous support, a combination of both, or a baseline system with no AI. We aimed for higher ecological validity than in typical AI-assisted decision-making studies. To this end, we validated and refined our scenarios in a focus group with four additional pilots prior to the study.

Our results challenge the common assumption that AI decision support should be recommendation-centric. Continuous support allowed pilots to think more beyond the limits of the system, was better accepted by pilots, and led to faster decisions when combined with recommendations. Our paper makes the following three contributions:

(1) We add to the rare examples of evaluative studies of AI-assisted decision-making with experts on a real-world task, in a domain that is understudied in the HCI community.
(2) We conduct one of the first empirical comparisons between recommendation-centric and alternative forms of AI decision support, demonstrating the importance and potential of thinking beyond the typical recommendation-centric paradigm.
(3) Based on our results, we propose a framework for *process-oriented decision support* as alternative to recommendation-centric support. We provide continuous support as well as further suggestions from our participants as concrete implementations of process-oriented support for diversions, but we consider the framework to be applicable in other domains beyond aviation as well.

2 Background and Related Work

We outline recent work on recommendation-centric decision support in Section 2.1 as well as alternatives to this dominant support paradigm in Section 2.2. We then describe the gap and the research questions we address in Section 2.3.

2.1 Recommendation-Centric Decision Support

The most common strategy to help decision makers work better with recommendation-centric AI is to add explanations of how the model works [2, 6, 39, 66, 74] and other model information,

such as model confidence [44, 52, 74] or the stated model accuracy [27, 52]. The goal is to help people to rely appropriately on AI recommendations [54] by giving cues about when they may be beneficial or detrimental to rely on. Results have been mixed so far, as especially explanations are prone to induce blind trust [2, 19] and hence overreliance, even among domain experts [6, 32]. Recently, Vasconcelos et al. [64] have shown that explanations can reduce overreliance under certain conditions, but it is questionable how often these conditions are valid in real applications [22]. Communicating model confidence appears more promising, as it has repeatedly been shown to improve appropriate reliance [2, 44, 48, 74]; but this relies on well-calibrated confidence scores, which are often difficult to achieve—models can be wrong with high confidence.

The limited success of adding model information appears to be due to people not engaging cognitively with it [8, 23]. One way to increase engagement is to employ *cognitive forcing* interventions, such as showing recommendations only after users made an initial decision [8, 21], introducing a waiting time until recommendations are shown [8, 46], or forcing users to wait before they can proceed to the next task [51]. While effective, these interventions negatively impact user experience [8, 21, 46]. Liu et al [43] explored the use of interactive explanations to increase engagement, though it did not reduce overreliance in their case.

Recently, an increasing number of voices call the entire premise of recommendation-centric support into question. Koon [36] emphasizes that decisions are often complex, and that condensing them into an AI recommendation is necessarily reductionist. At the same time, since recommendations are hard to appropriate, users often struggle to combine them with the wider context knowledge they have [9]. From a cognitive science perspective, Miller [45] argues that recommendation-centric support does not align with the cognitive processes of human decision-making. Instead, recommendations take control away from decision makers. Similarly, Wang et al. [65] and Zhang et al. [75] caution against error-prone *backward reasoning* from the end result back to the input data, which is facilitated by a fixation on end-to-end recommendations. All of these authors call for alternative approaches to AI decision support that are less centered on recommendations.

2.2 Alternative Forms of AI Decision Support

One stream of work that de-emphasizes recommendations rethinks the purpose of explanations. Instead of explaining the AI model, explanations can provide information that is of natural interest in a decision, such as domain-specific information [41, 69], or the socio-organizational context of a decision [18]. These explanations situate recommendations within the primary decision-making task, rather than diverting attention to the secondary task of understanding the AI model.

Other authors propose entirely different roles for AI than providing end-to-end recommendations. Alternative AI support paradigms are far from new [67], but have been largely ignored by current research. This is arguably more due to technical feasibility rather than consideration for human needs, since with modern AI methods, it is straightforward to formulate many decision tasks as end-to-end predictions [7, 58]. As for alternatives to providing recommendations, on a conceptual level, Cabitza et al. [9] propose to frame AI as "knowledge artifact functions" which support people in their collaborative decision-making. Zhang et al. [75] put forward what they call "forward-reasoning decision support", where users form decisions themselves, augmented by rich interactions with AI tools. In a similar, but more concrete way, Miller [45] proposes the concept of "evaluative AI", which focuses on helping decision makers to evaluate different hypotheses.

In essence, all of these proposals aim to help people to make decisions through *forward reasoning*, allowing people to start from the context at hand and to use their domain expertise to reach a decision. This is in contrast to recommendation-centric support, which pushes people to reason backward from the recommendation. In fact, cognitive forcing can be seen as a way to encourage forward reasoning as well by pushing people to think independently from AI recommendations.

However, with cognitive forcing interventions, people get no support while making their independent decisions, which reduces the supportive value of the AI [8, 21, 46]. The above concepts aim to facilitate forward reasoning while supporting users' decision-making processes.

Beyond abstract concepts, concrete examples for supporting decisions without relying on end-to-end recommendations are often found in healthcare. Lindvall et al. [42] designed a system for tumor assessment that navigates pathologists to potentially tumorous image regions for them to review, without revealing whether or how confidently the model classifies the pixels as tumor. Crucially, the classification threshold was not set to optimize accuracy, but sensitivity. Consequently, if pathologists did not find a tumor in any of the suggested regions, they could be relatively sure that the rest of the image also contains no cancer. Zhang et al. [73] studied the case of sepsis diagnosis, where an existing AI tool only addresses the final stage of the decision-making by offering a sepsis risk score and sending alerts above a certain threshold. The authors proposed a redesign where the system suggests lab tests that would help to reduce uncertainty about patients' future conditions.

In the context of aviation, Zhang et al. explored how to support diversions decisions [76]. Their system does not recommend airports, but continuously provides unobtrusive local hints about potential limitations at the surrounding airports. The system provides this support of pilots' SA also during normal flight when there is no sign of an emergency yet.

2.3 Summary and Research Questions

There is a remarkable difference between the studies mentioned in the previous two sections: The studies on recommendation-centric support in Section 2.1 were almost all built on simple, artificial tasks with lay persons as participants. Such studies make up the majority of research in AI-assisted decision-making [38]. In contrast, the alternative support paradigms in Section 2.2 all stemmed from studying complex real-world decisions with domain experts. This indicates a potentially significant gap between what is studied in large-scale controlled experiments and what is actually required by experts in real applications. Our study is situated right in this gap by conducting a controlled comparison between alternative decision support paradigms with domain experts.

The goal of our work is to empirically compare continuous support with typical recommendation-centric support. We further add a combination of both to the comparison, where the system provides continuous support during normal flight, but gives recommendations in an emergency. Figure 1 contrasts how the two paradigms conceptually fit into pilots' decision-making, based on the *FOR-DEC* model [30]. FOR-DEC is a prescriptive model used by many airlines to train pilots to make decisions in a structured manner. It is an acronym for the steps to follow during decision-making: *facts*, *options*, *risks & benefits*, *decision*, *execution*, *check*. The dash in the middle signals a pause for reflection before making the decision. Figure 1 only covers the first four steps of the model, since execution and check are beyond the scope of our work.

The focus of our empirical comparison is to assess whether pilots reason differently with continuous support and recommendation-centric support, and how this affects pilots' perceptions about the system as well as decision outcomes, namely overreliance and decision time. As outlined in Section 2.1, the former is one of the major concerns in recommendation-centric support. The latter is of interest since aviation is highly dynamic, which limits the time available for decision-making. We therefore pose the following research questions:

- **RQ1**: How do pilots integrate the different support paradigms into their workflow?
- **RQ2**: How do the support paradigms differ in terms of overreliance?
- **RQ3**: How do the support paradigms differ in terms of decision time?
- **RQ4**: How do pilots perceive the different support paradigms?

We hypothesize that recommendations induce backward reasoning, while continuous support facilitates forward reasoning. We further hypothesize that combining both allows to give recommendations while still enabling forward reasoning. Hence, we expect continuous support to lead to less overreliance than recommendations, but at the expense of longer decision times. We expect the combination of both to combine their advantages, i.e., less overreliance and shorter decision times.

Most closely related to our work is a study by Smith et al. [60], who compared the effects of three different versions of a flight planning tool on users' performance. Apart from being a different task, the major difference to our work is that their system versions represent consecutively increasing levels of automation, while we study alternative support paradigms.

3 Diversion Assistance System

In this section, we present the design of the four versions our diversion assistance system (Section 3.1), followed by the apparatus used to conduct our study (Section 3.2).

3.1 System Design and Variants

Our diversion assistance system (DAS) consists of three basic components (⑥ in Figure 3). A navigation map in the top left of the main screen shows surrounding airports relative to the current aircraft position. The bottom half of the main screen shows the same airports in a table, displaying runway, weather, and operational information as well as the time to and fuel remaining at each airport. The table can be sorted by any of the columns by tapping the respective column header. The details box to the right of the navigation map shows additional raw information for the selected airport, including raw runway data, weather reports, and NOTAMs[1].

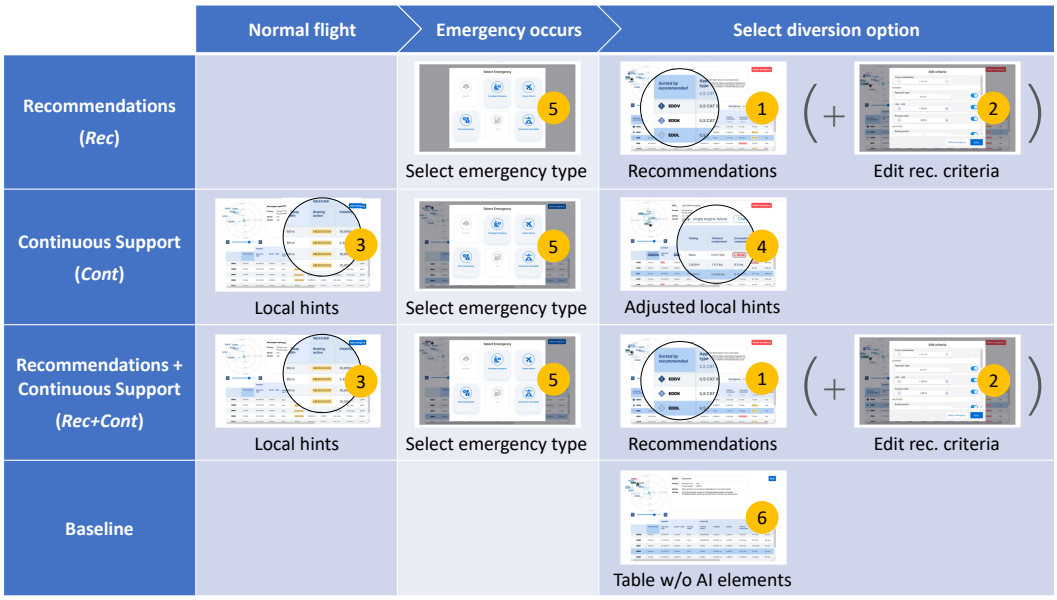


Fig. 2. The four versions of the diversion assistance system, from top to bottom: *Rec*, *Cont*, *Rec+Cont*, *Baseline*. The corresponding views ①–⑥ that are applicable during the three operational phases *normal flight*, *emergency occurs*, and *select diversion option*, are numbered corresponding to the detailed view in Figure 3.

[1] *Notice to Air Missions*, formerly *Notice to Airmen*. These are messages containing information about abnormal conditions that may affect flight operations, e.g. runway closures, inoperable systems at an airport, or drones near the runway.

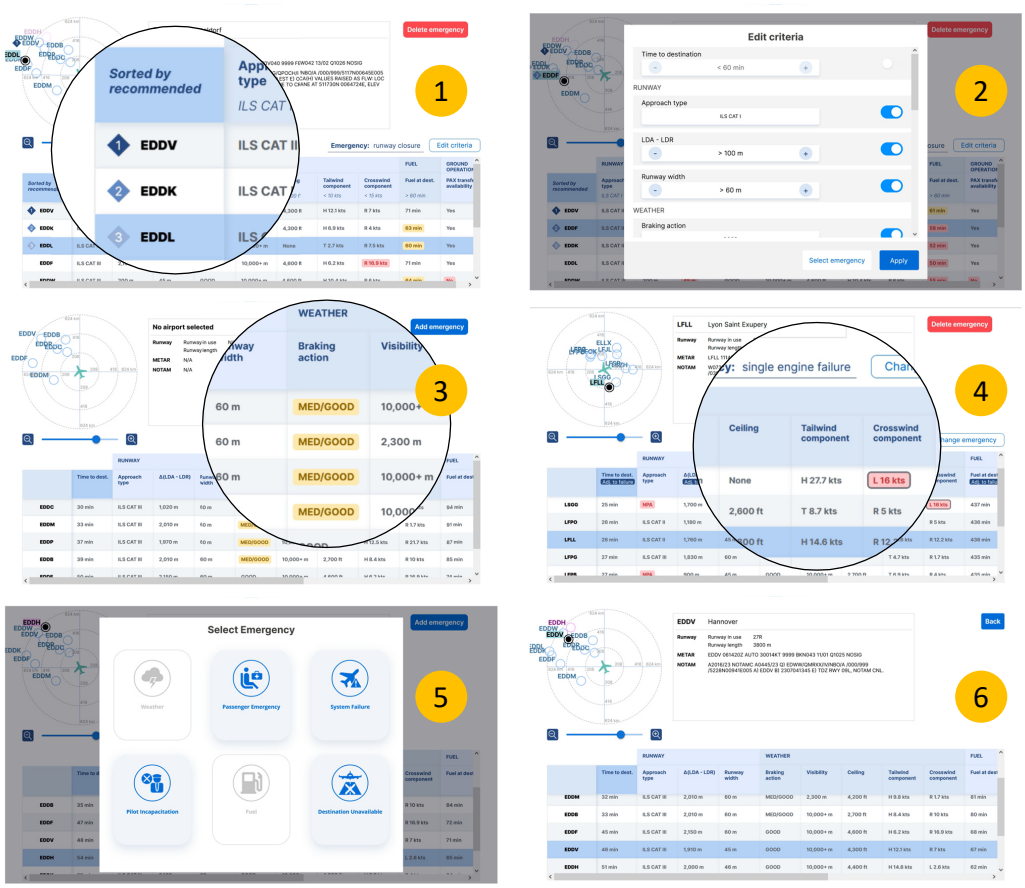


Fig. 3. Screenshots of the diversion assistance system. ① Recommendations. ② Edit recommendation criteria. ③ Local hints during normal flight. ④ Adjusted local hints during emergency. ⑤ Select emergency type. ⑥ Baseline without AI. The numbers correspond to those in Figure 2.

Based on these basic components, we designed four versions of the DAS, as shown in Figure 2. In the **Recommendations** version, the system sorts the table by the AI's evaluation of each airport and recommends up to three options, which are marked in both the table and the navigation map (① in Figure 3). The AI's evaluations are based on pre-defined criteria, like having a certain amount of fuel left at an airport. These criteria are shown in their respective column header and can be edited by pilots if desired (② in Figure 3). Yellow and red highlights indicate when a criterion is only barely met or not met at an airport, providing transparency about the AI's evaluation. The table can be sorted manually by each criterion by tapping the respective column header. The intended flow of this system version is as follows: When an emergency or abnormal situation happens, pilots first select the emergency type (⑤ in Figure 3), after which the main screen with the recommendations shows up. The pre-set evaluation criteria are catered to the selected emergency type. This system variant represents the common recommendation-centric paradigm of AI decision support.

In contrast, the **Continuous Support** version shows the main screen also in normal flight when there is no emergency in sight yet. In this normal flight mode, the system continuously evaluates the surrounding airports for potential constraints that are of general interest for any type of diversion,

like a wet runway (③ in Figure 3). These potential constraints are shown in the table, again as yellow and red highlights. Instead of explaining recommendations—which this system variant does not have—the highlights serve as local hints to guide pilots' attention. This normal flight mode is meant to support pilots' SA and their continued preparation for hypothetical emergencies.

In an emergency, pilots can switch to emergency mode, again by selecting the emergency type, which adjusts the hints to the selected emergency (④ in Figure 3). For instance, 16 knots of crosswind may not be a big concern normally, but with an engine failure, it might be critical. The system would highlight the crosswind, which it did not highlight in normal flight. New hints are indicated by a solid black border. At no point does this system variant generate decision recommendations. The table is always sorted by the pilot-selected column, and per default by *time to destination*.

The **Recommendations + Continuous Support** version combines the two prior versions, with continuous support in normal flight, and recommendations in case of an emergency. Lastly, we added a **Baseline** version, which only shows the main screen with neither recommendations nor local hints (⑥ in Figure 3) and which is only available in case of emergency.

3.2 Apparatus

As simulation environment, we chose X-Plane 11[2], running the Airbus A320 Ultimate aircraft model by Flight Factor[3]. We supplemented the simulator with hardware controls, including a sidestick and a throttle quadrant from Thrustmaster[4], as well as an MCDU[5] and an FCU[6] from Skalarki[7], to reduce reliance on a mouse for interacting with the aircraft. The DAS was implemented as an interactive mockup in the prototyping tool Framer[8], with custom components written in React. It runs on a second-generation Surface Pro tablet next to the simulator computer, mimicking an EFB[9]. The mockup communicates via Websocket with a Python back end on the simulator computer, which acts as an interface to X-Plane to read out live flight data. Weather and ground facilities at each airport are hard-coded for each scenario. The Python back end further triggers the emergencies of the study scenarios in X-Plane based on timers. The entire setup is shown in Figure 4. Note that despite the remarkable realism of the elements we used in our simulation setup, some notable deviations from real flights remained. We describe these deviations and discuss their implications in Section 6.4.

The AI recommendations are generated with a manually-tuned scoring function that assigns each airport a score, calculated as a weighted sum of subscores for each criterion. Each subscore is calculated using a piecewise linear function that penalizes unfulfilled criteria more heavily than it rewards overfulfilled criteria. Using this scoring function, the recommendations react dynamically to what is happening in the simulator and to the criteria the pilot defines. While it is not machine learning, given the large number of criteria, it is still hard to comprehend what exactly causes a certain airport to be recommended or not, just as with an opaque deep neural network.

4 Methods

Besides the DAS, valid and informative scenarios were the key component of our study. To ensure the quality of our scenarios, we first conducted a focus group to discuss and flesh out our initial

[2]https://www.x-plane.com/
[3]https://flightfactor.aero/
[4]https://www.thrustmaster.com/
[5]*Multipurpose Control and Display Unit*, the input device for the flight management system and other computer systems.
[6]*Flight Control Unit*, the control panel used to control the autopilot.
[7]https://www.skalarki-electronics.com/
[8]https://www.framer.com/
[9]*Electronic Flight Bag*, a portable device to store and display flight-relevant data and to run assistive applications.

Fig. 4. Overview of the study apparatus.

scenario outlines with professional pilots. We subsequently tested our scenarios and system design with pilots before finally running our main study. In this section, we describe these steps and our methodology in detail, which are shown in Figure 5.

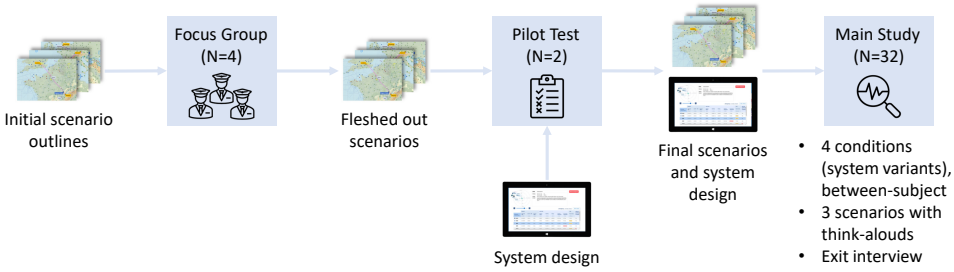


Fig. 5. Overview of the study parts.

4.1 Focus Group and Study Scenarios

There is no inherent right or wrong in most diversion decisions, so one goal of the focus group was to obtain a reference for how pilots would decide in each scenario and for which reasons. The other goal was to understand how to design the details of the scenarios to fit our intentions. We recruited four professional pilots for the focus group (one captain, three first officers; all male; median age: 29.5 years (IQR 28.5–34.75); median flight hours: 1000 hours (IQR 392.5–3750); details in Appendix B, Table 3). All four work at the same airline, but have past experience in four additional airlines. We showed them the initial outline of each scenario and asked them to discuss how they would decide. The focus group had a duration of 90 minutes, and participants were compensated with 150 EUR (≈ 164 USD) each, which is a typical rate for professional pilots, given the difficulty to recruit them.

After the focus group, we adjusted the scenario outlines and filled in the details according to the insights gained. In total, we designed three scenarios, with the aim to cover a good range of common

reasons for diversions as well as different potential failure modes of the DAS, as recommended by Roth et al. [53]. In particular, we aimed to construct one scenario each where the DAS performs well (Scenario 1), suggests an airport that pilots tend to disagree with (Scenario 2), and suggests a solution that is subpar, but for a reason that is not immediately obvious (Scenario 3).

Fig. 6. Single engine failure scenario. The flight is from Geneva to London. Figure based on SkyVector[10].

4.1.1 Scenario 1: Single engine failure. In the first scenario, a flight from Geneva, Switzerland to London Heathrow suffers a single engine failure during climb (Figure 6). The scenario is based on incident reports found with the CAROL query tool of the National Transportation Safety Board[11]. The focus group participants pointed out that the terrain around Geneva may be challenging with an engine failure, but returning to Geneva is likely preferable for the company. Charles de Gaulle on the other hand is a good option since it is a major hub with excellent infrastructure.

Our intention for this scenario was that the DAS would recommend Charles de Gaulle as an option that pilots would agree with. Beginning with this scenario was meant to establish pilots' initial trust into the system, since in reality, such a system would perform well in most situations. To make the decision less ambiguous and hence the recommendation more acceptable, we added a moderately high crosswind component to Geneva. In the scenario, we triggered an engine failure and displayed a popup message within X-Plane, reading *"Engine 1 has failed. No engine relight*[12]*. Please secure the engine and proceed with the diversion decision."* Pilots were asked to handle the engine failure within X-Plane as they would in reality to keep them in their known workflow.

4.1.2 Scenario 2: Passenger medical emergency. The second scenario was a flight from Tenerife, Spain to Munich, Germany, with a passenger having a heart attack while the flight is above the sea in the middle between Tenerife and the European continent (Figure 7). The emergency was announced through a popup message in X-Plane: *"The passenger on 11C has vomited, complains of extreme chest pains, is pale and sweating. This has been going on for a few minutes now. A doctor (general practitioner) sitting next to the passenger reacted immediately. He suspects a heart attack."*

[10]https://skyvector.com/
[11]https://carol.ntsb.gov/
[12]In reality, pilots would try to relight the engine. We asked them to skip this to simplify the scenario, as it would not have added any value to our study.

Fig. 7. Medical emergency scenario. The flight is from Tenerife to Munich. Figure based on SkyVector.

This scenario was based on a route that is notorious among short-haul flight pilots due to the long flight over open sea, with much fewer diversion options than usual on European short-haul flights.

Our expectation was that Casablanca, Morocco would be the obvious option since it is the closest, making the decision too easy. Hence, we thought about adding a reason to make Casablanca unattractive, like a nation-wide healthcare strike. However, our focus group participants said they would either fly to Faro, Portugal, or return to Tenerife, depending on which is faster, since they would prefer to stay within Europe. This would make it easier to continue the flight and was also assumed to allow for better medical treatment for the passenger. We therefore decided not to add any factor to make Casablanca unattractive. Still, the DAS would recommend Casablanca to probe for pilots' reactions when the system recommends an option that they are not comfortable with.

Fig. 8. Airport closure scenario. The flight is from Sharm El Sheikh to Hamburg. Figure based on SkyVector.

4.1.3 Scenario 3: Airport closure. This last scenario was based on the personal experience of one participant in Zhang et al.'s study [76]. The flight was from Sharm El Sheikh, Egypt, to Hamburg, Germany (Figure 8). About an hour before landing, participants were given the information that

the destination had closed down, again via a popup message: *"Due to a power outage, EDDH*[13] *had to close completely. It will remain closed for at least the next few hours."*

The intention for this scenario was that the DAS would recommend an option that is good at first glance, but problematic at second thought, in order to probe for how much participants would over-rely. The system would recommend Hanover, Germany, which is the typical alternate for Hamburg. It also has good ground connections to Hamburg for the passengers and is therefore a plausible recommendation. However, the airport closure affects the entire air traffic coming into Hamburg. Many planes would try to divert to Hanover, which in the real event had caused Hanover to run full. The DAS does not consider this traffic factor.

Unprompted, focus group participants discussed the traffic situation in Hanover as an important consideration, confirming that pilots would realistically think about it in such a situation. Participants said that instead of Hanover, all the en-route airports like Berlin, Dresden, or Leipzig would be valid options. We added slightly unfavorable, but workable weather conditions at these airports, to make the system recommendation Hanover seem more attractive at first glance.

4.2 Main Study

The main study had a between-subject design where each participant was assigned to one of the four DAS variants in Figure 2 and completed all three scenarios in the order of Section 4.1. At the beginning of the study, pilots got a detailed introduction into X-Plane and the DAS. Before starting each scenario, pilots were shown maps similar to those in Figure 6–Figure 8, with the flight plan, the departure and destination airport, as well as the position of the aircraft at the beginning of the scenario. The scenarios started a couple minutes before the incident to allow participants to familiarize themselves with the situation. Participants were allowed to freely interact with the flight simulation, but were asked to only call up the DAS (with *Recommendations* and *Baseline*) or enter emergency mode (with *Continuous Support* and *Recommendations + Continuous Support*) when the incident happened. With *Continuous Support* and *Recommendations + Continuous Support*, pilots were allowed to freely interact with the normal flight mode before the incident.

Participants were asked to think aloud and clearly announce which airport they would divert to; they did not need to execute the diversion. The pilots were encouraged to think beyond what they saw in the DAS and consider whatever they would in a real flight, including which other stakeholders they would contact, like their company or air traffic control. They were further told that whatever is not shown in the table is not considered by the AI.

After the scenarios, we conducted semi-structured exit interviews with each participant to discuss their impressions of the system and how they used it. The interview guide is given in Appendix A. Each study session took around 90 minutes, for which participants were paid 150 EUR (≈ 164 USD) each. Again, this is a typical rate for pilots due to the difficulty of recruiting them. We recorded audio as well as the screen of the tablet running the DAS. The study was approved by the Ethics Committee of the Faculty of Mathematics, Computer Science and Statistics at LMU Munich.

We pilot-tested the study procedure with two of the focus group participants from Section 4.1 (see Appendix B, Table 3), whom we also paid 150 EUR each. Following the pilot test, we made small adjustments to the DAS user interface and to the third scenario. Most importantly, both pilot testers asked for preferences of the company in Scenario 3. We therefore decided that if participants would ask for this information, we would tell them that the airline would prefer a diversion to either Hanover, Berlin, or Düsseldorf. Hanover was the top recommendation of the AI, Düsseldorf was the third recommendation, and Berlin was not recommended due to slightly worse weather.

[13]The ICAO (International Civil Aviation Organization) airport code for Hamburg.

4.3 Data Analysis

We took participants' decisions and the time they needed for their decisions as outcome measures. As decision time, we took the time between calling up the system or entering emergency mode and announcing the decision. We further noted the reasons for the decisions from the think-aloud protocols to cover both outcome and process measures, as recommended by Roth et al. [53].

To analyze the qualitative data, we transcribed the think-aloud protocols and exit interviews and coded both parts through thematic analysis [3]. An initial round of open coding was conducted independently by two authors on four transcripts, one per study condition. The two authors then discussed their initial set of 269 low-level codes and consolidated them into 18 code groups. The first author coded the rest of the transcripts, extending and revising the initial coding scheme when necessary. All additional and revised low-level codes in this step continued fitting into the 18 code groups. Finally, we identified three themes that reflect the code groups.

Additionally, we extracted usage patterns from the think-aloud protocols by reviewing the codes in their temporal order in each of the 96 decision instances (32 participants × 3 scenarios), referring back to the screen recordings where necessary for more context.

5 Results

For our main study, we recruited 32 professional pilots through snowball sampling (two captains, rest first officers; two female, rest male; median age: 31 years (IQR 30–35); median flight hours: 2500 hours (IQR 2000–4000); median of 1.5 self-performed diversions (IQR 0–3); details in Appendix B, Table 4). The pilots work in four German airlines, with past experience in eight additional airlines.

We structure our results according to our research questions from Section 2.3. Hereafter, we use the abbreviations *Rec*, *Cont*, and *Rec+Cont* according to Figure 2 for the respective system variants. Participants are denoted with Rx, Cx, RCx, and Bx according to the system variant they used.

5.1 RQ1: Workflow Integration

We first describe the usage patterns we identified, followed by a comparison of how often they occurred across scenarios and system variants.

5.1.1 Usage patterns. All participants intuitively integrated the DAS into the FOR-DEC framework they are familiar with. We therefore distinguish the identified usage patterns on two overarching levels: the options considered by participants, and the strategies they used to decide between the options. The first directly mirrors the *options* step in FOR-DEC, while the latter is a combination of the *risks & benefits* and the *decision* steps, which were not always clearly distinguishable in the think-aloud protocols. We identified three distinct patterns for the options considered:

- **O1—First few options**: Participants considered the first few options in the table. For *Rec* and *Rec+Cont*, these were the AI recommendations. For *Cont* and *Baseline*, these were the closest airports, since the table was sorted by time per default.
- **O2—First option**: Like the above, but participants only considered the very first instead of the first few options in the table.
- **O3—Self-generated options**: Participants generated their options independently from the order in the table, e.g. by sorting the table according to a column of interest, by looking for familiar airports, or by asking for company preferences.

We further identified twelve strategies that participants used to decide between the options they considered. The first five are general strategies that could be used across all variants:

- **S1—Narrow down**: Participants carefully compared the considered options, ruling out options one by one until one remains.

- **S2—Recognize best**: Participants immediately recognized the best among the considered options and focused on it without detailed comparison against the other options.
- **S3—Check if first option works**: In case participants considered only the first option, they checked whether there was anything against it.
- **S4—Confirm first option is best**: In case participants considered only the first option, they did a quick cross check to confirm that it was indeed the best, e.g. by glancing over the *time to destination* to see if the first option was significantly closer than the next ones.
- **S5—Check one after another**: Participants checked whether there was anything against the first option. If yes, they checked the second one, and so on.

The seven remaining strategies capture how participants used the AI support elements available to them. The normal flight mode of the *Cont* and *Rec+Cont* variants was used in two different ways:

- **S6—Use prepared plan**: Participants prepared a plan for a hypothetical emergency during normal flight. Even though they did not know what would happen, they could prepare on the level of e.g., *"If something very urgent would happen, x would be a good option."* When the emergency happened, participants used this plan to quickly reach a decision.
- **S7—Refine situation awareness**: During normal flight, participants were aware of the general situation, such as the weather around them. When the emergency happened, participants did not review this information again, but only supplemented it with situation-specific information like the distance to the next hospital in Scenario 2.

The color highlights, which served as transparency for recommendations and as local hints for continuous support, were also used in two ways:

- **S8—Look for options without highlights**: Participants reviewed highlights for their relevance, but gravitated toward options without highlights. This strategy was used for choosing among the considered options as well as for deciding which options to consider in the first place. For instance, participants looked at *time to destination* and quickly excluded those options with red highlights, which indicated that these airports were too far away.
- **S9—No highlights as confirmation**: Participants had a decision in mind and felt confirmed if that option had no highlights. The difference to above is that there, the highlights were used as cue to guide the decision-making, while here, participants considered the highlights as a second opinion to their own independent reasoning.

Lastly, participants employed three distinct strategies for using recommendations:

- **S10—Recommendation as confirmation**: Participants had a decision in mind and felt confirmed if that option was also recommended by the DAS. Some pilots using this strategy explicitly ignored the recommendations initially to reach a decision on their own first.
- **S11—Recommendation as fallback**: Participants first checked their self-generated options. When they were not satisfied with any of them, they chose the AI-recommended option.
- **S12—Negotiate**: Participants edited the criteria to see how it affected the recommendations. This strategy was triggered by one of two reasons, or a combination of both: Some pilots disagreed with the pre-defined criteria, while others were surprised that their favored option was not recommended. In the latter case, they tried to align the recommendations with their own opinion by removing or relaxing criteria they deemed uncritical. Those that successfully aligned the recommendations to their favored option took this as confirmation. One pilot did not succeed, which triggered further considerations in his decision-making.

5.1.2 Differences between scenarios and support paradigms. Figure 9 gives an overview of the occurrences of each usage pattern across conditions and scenarios. Note that the real occurrences could be higher, since participants might have used certain strategies without any revealing

		OPTIONS CONSIDERED			DECISION STRATEGY											
					General strategies					AI usage strategies						
										Normal flight mode		Highlights		Recommendations		
		First few options (O1)	First option (O2)	Self-generated options (O3)	Narrow down (S1)	Recognize best (S2)	Check if first o. works (S3)	Confirm first o. is best (S4)	Check one after another (S5)	Use prepared plan (S6)	Refine SA (S7)	Look for o. w/o highlights (S8)	No highlights as conf. (S9)	Rec. as conf. (S10)	Rec. as fallback (S11)	Negotiate (S12)
SCENARIO 1 Engine failure	Baseline	5		4	2	5			1							
	Rec	8		2	3	5						2	1	1		
	Cont	3		6	3	6		1		2		1	1			
	Rec+Cont	5	3	3	4	3	2									
SCENARIO 2 Passenger emergency	Baseline	4	4		2	2	2	2								
	Rec	2	4	2	3	1	4	4						2		2
	Cont	1	6	1		2	3	2		5	3	1	1			
	Rec+Cont	2	6		1		2	3		3				6		
SCENARIO 3 Airport closure	Baseline	1		7	5	3										
	Rec	5		8	5	2								1	2	2
	Cont	1		8	5	3				2		3				
	Rec+Cont	6	1	8	4	4						1	1	2	1	1

Fig. 9. Usage patterns across conditions and scenarios. The counts exceed the number of participants because individual usage patterns are not mutually exclusive and participants typically employed more than one pattern. Grayed-out fields mean that the patterns are not applicable to the respective system variant.

verbalizations. Especially some of the AI usage strategies were more subconscious and therefore less likely to be verbalized.

The most obvious difference is between the options considered in different scenarios, where participants tended to consider the *first few options* (O1) in Scenario 1 and only the *first option* (O2) in Scenario 2, while they based their decision more heavily on *self-generated options* (O3) in Scenario 3. The difference was mostly due to the varying time criticality, as several participants explained during the exit interviews. The engine failure in Scenario 1 was of medium time criticality, the passenger emergency in Scenario 2 was extremely time-critical, while the airport closure in Scenario 3 was not time-critical at all.

Besides these scenario differences, there are also differences between the DAS variants. These are most salient in Scenario 2, where participants in the *Cont* and *Rec+Cont* groups benefitted the most from the possibility of detailed pre-planning with the normal flight mode. As a result, pilots in these groups could *use their prepared plan* (S6) or use *recommendations as confirmation* (S10) for their plan, indicating forward reasoning by these participants. Note that pilots using the *Baseline* and *Rec* variants also tried to establish SA during normal flight, as they would in reality, but could only rely on the much more limited possibilities offered by X-Plane. Consequently, when the emergency happened, pilots in the *Rec* group were more likely to take the system recommendation as starting point for review (S3, S4), which indicates backward reasoning. In the absence of AI support that helps to identify the best option, participants using the *Baseline* variant had a stronger tendency to review the *first few options* (O1) rather than only the *first option* (O2).

In Scenarios 1 and 3, the benefits of pre-planning during normal flight in *Cont* and *Rec+Cont* are less apparent. This can be explained by a disruption between normal flight and emergency decision-making in both scenarios. In Scenario 1, the disruption happened because pilots first handled the engine failure before entering the diversion decision, a procedure that took several minutes during which the position of the aircraft changed significantly. After securing the engine, pilots therefore had to slightly re-orient and could not directly use their SA from the normal flight. In Scenario 3, the disruption occurred since pilots mostly prepared for a time-critical emergency requiring a quick landing. What happened instead was a situation where it was more important to find a suitable airport near the destination so that participants' preparations were not applicable.

Nevertheless, a difference is still observable between *Rec* and *Cont* in Scenario 1, indicating that continuous support encourages forward reasoning, while recommendations prompt backward reasoning. All participants using the *Rec* variant took the system recommendations as starting

point (O1), while participants in the *Cont* group tended to consider *self-generated options* (O3), suggesting that they could build on their normal flight SA to some extent, despite the disruption of handling the engine failure. Furthermore, some participants in the *Rec+Cont* group limited their options to the system's top recommendation (O2), which could have been because it conformed to their impressions of the available options from the normal flight phase.

In Scenario 3, a stark contrast is apparent between *Baseline* and *Cont* on the one hand—which did not have AI recommendations—and *Rec* and *Rec+Cont* on the other hand. Almost all participants in this scenario considered *self-generated options* (O3), mostly by asking for company preferences and by looking for familiar airports. In addition, those participants who had recommendations available to them also mostly considered them (O1), indicating a strong influence of the recommendations. However, from the think-aloud protocols, we were not able to tell whether they influenced participants' reasoning differently between *Rec* and *Rec+Cont*, given that the observable recommendations usage patterns are very similar (S10, S11, S12). Still, the observed usage patterns presented in Figure 9 suggest that our hypothesis holds true according to which recommendations prompt backward reasoning, while continuous support facilitates forward reasoning, even though the latter can be derailed by disruptions between normal flight and emergency decision-making.

5.2 RQ2: Overreliance

Decision outcomes were quite uniform for the first two scenarios. For Scenario 1, this was to be expected, since the scenario was designed to be rather unambiguous for the reason given in Section 4.1.1. All participants except for C5 (Lyon) and B8 (Paris Orly) decided for Paris Charles de Gaulle in Scenario 1, which was also the system's top recommendation. However, in Scenario 2, all participants except for B5 (Faro) chose Casablanca, which ran contrary to our expectation given the discussion of our focus group participants, as described in Section 4.1.2. Some participants in the main study, especially those who had flown in the region themselves before, emphasized that the bias against diverting to Africa is common among their colleagues, but unwarranted in this case. 14 participants across all conditions did mention that they would generally prefer to divert to Faro for the reasons also discussed by the focus group participants. In the end, the fact that the emergency was urgent and that Casablanca was around ten minutes closer tipped the scales for all of these pilots. The AI elements—which favored Casablanca—did not seem to have an effect on this decision, given that almost all participants in the *Baseline* group also decided for Casablanca.

As intended, Scenario 3 was less clear for participants, as shown in Table 1. Of most interest in terms of overreliance was how many participants decided for Hanover without considering the traffic situation, as explained in Section 4.1.3. While some participants did rule out Hanover themselves because of the traffic, other pilots either said they would ask air traffic control about the traffic density, or they considered traffic but expected it to be not too dense to land there. We told these participants that Hanover was already running out of capacity, as air traffic control would do in reality. We did so since we were not interested in how pilots would assess the traffic situation, but only whether they would consider it at all. The DAS does not include this information, so we were interested in whether this would lure pilots into overlooking this factor, or whether they would think beyond the limits of the system.

Figure 10 shows how the three AI system variants affected the probability for the overreliant behavior of choosing Hanover without considering traffic, as compared to the *Baseline* variant. Since we assumed that adding AI would increase the probability for overreliance, we performed one-sided Fisher's exact tests to compare each AI variant with *Baseline*. We further report relative risks (RR) compared to *Baseline* with Wald normal approximation confidence intervals as effect sizes. Consistent with our hypothesis, there was a statistically significant increase of overreliance probability for *Rec* ($RR = 6$, 95% CI [0.92, 39.18], $p = 0.02 < 0.05$). For *Cont* ($RR = 4$, 95% CI [0.56,

Table 1. Diversion decisions across conditions in Scenario 3. Hanover was to be avoided due to heavy traffic, which was not considered by the system.

		Baseline	Rec	Cont	Rec+Cont
To be avoided	**Hanover** *AI recommendation & company preference*	1/8	6/8	4/8	3/8
Other	**Berlin** *company preference*	7/8	-	4/8	3/8
	Düsseldorf *company preference*	-	1/8	-	1/8
	Cologne	-	-	-	1/8
	Frankfurt	-	1/8	-	-

28.40], $p = 0.14$) and *Rec+Cont* ($RR = 3$, 95% CI [0.39, 23.07], $p = 0.28$), the increase was smaller and not statistically significant, though these latter results have to be interpreted carefully given the small sample sizes and the resulting large confidence intervals. However, it appears like even without recommendations, pilots using the *Cont* variant could still be biased by the hints: Some airports in Scenario 3, including Berlin, had a braking action of "medium to good", i.e., slightly wet runways, which the system highlighted in red. With the *Baseline* variant, which did not have highlights, participants acknowledged the wet runways with a short comment that it was uncritical; some even did not verbalize any thoughts about it at all. By contrast, with the AI variants, many participants found the red highlights hard to ignore, even though they deemed the braking action uncritical: *"To me, red is always intuitively, it doesn't look so good. That's why I looked a bit at Hanover, even though 'medium to good' is fine."* (C8). This behavior is also reflected by the usage patterns in Figure 9, where some *Cont* participants looked for options without highlights (S8).

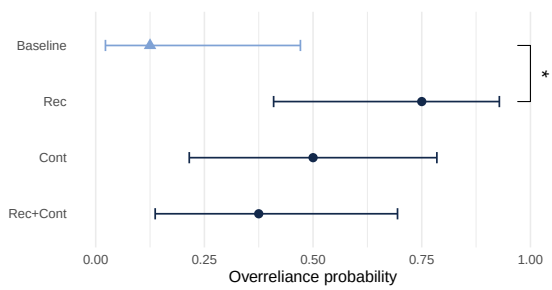


Fig. 10. Probability of overreliance across conditions, where choosing Hanover in Scenario 3 without considering traffic conditions is counted as overreliance. Error bars denote 95% confidence intervals, estimated using the Wilson score interval method. (*) denotes $p < 0.05$.

In addition to the decisions, we also noted the reasons behind them. Table 2 gives an overview of the reasons why non-overreliant participants decided against Hanover, showing that traffic density was indeed the reason for most of them. Some participants decided against Hanover but not due to traffic. These pilots—all of whom saw no recommendations—favored Berlin for other reasons and focused on it without discussing Hanover in depth. Interestingly, the pilots in the *Rec* group who thought of traffic were two of the three pilots in that group who ignored the recommendations (i.e.,

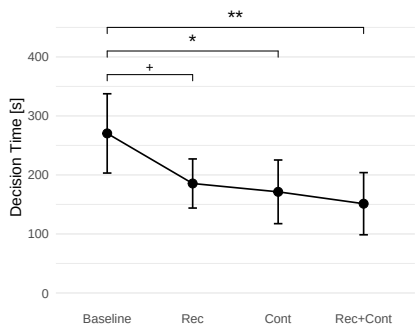
not used the pattern O1 in Figure 9) to make their own independent thoughts. This suggests that forward reasoning was important for pilots to be able to think beyond the limits of the system.

Table 2. Reasons for not diverting to Hanover in Scenario 3. Participants could give more than one reason.

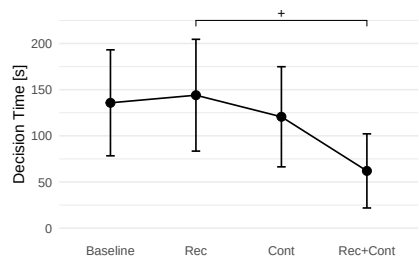
	Baseline	**Rec**	**Cont**	**Rec+Cont**
Traffic in Hanover	5/7	2/2	3/4	5/5
Berlin is closer to current location	2/7	-	1/4	-
Berlin has better connections for crew to return home	-	-	1/4	-

5.3 RQ3: Decision Time

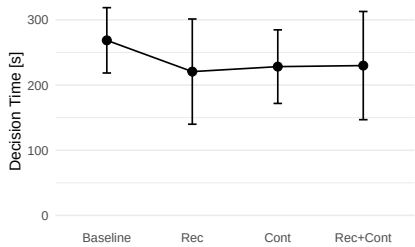
The differences in usage patterns discussed in Section 5.1.2 also resulted in different patterns in decision times, as shown in Figure 11. We therefore analyzed each scenario separately through an analysis of variance (ANOVA). We further used Tukey's HSD test for post-hoc tests and report both mean differences in seconds and Cohen's d as effect sizes.



(a) Scenario 1 (engine failure)



(b) Scenario 2 (passenger emergency)



(c) Scenario 3 (airport closure)

Fig. 11. Decision time across conditions for all three scenarios. Error bars represent 95% confidence intervals. (+) denotes $p < 0.1$, (*) denotes $p < 0.05$, and (**) denotes $p < 0.01$.

In Scenario 1, we found statistically significant differences in decision times at the $p < 0.01$ level ($F(3, 28) = 5.15, p = 0.006$), with a large effect size of $\omega^2 = 0.28$. Post-hoc tests revealed that all

three AI system variants led to significantly faster decisions than *Baseline*: *Rec* was significantly faster than *Baseline* at the $p < 0.1$ level ($\Delta = 84.88$ s, 95% CI [-4.30 s, 174.05 s], $d = 1.30$, 95% CI [-0.20, 2.80], $p = 0.066$), *Cont* at the $p < 0.05$ level ($\Delta = 99.00$ s, 95% CI [9.83 s, 188.17 s], $d = 1.52$, 95% CI [-0.02, 2.05], $p = 0.025$), and *Rec+Cont* at the $p < 0.01$ level ($\Delta = 119.13$ s, 95% CI [29.95 s, 208.30 s], $d = 1.82$, 95% CI [0.25, 3.40], $p = 0.006$). The effect sizes can be considered large in all three comparisons. The differences between the AI variants were not statistically significant.

In Scenario 2, decision times differed significantly at the $p < 0.1$ level ($F(3, 28) = 2.66, p = 0.067$) with a medium effect size of $\omega^2 = 0.135$. Post-hoc test results showed that *Rec+Cont* was significantly faster than *Rec* at the $p < 0.1$ level with a large effect size ($\Delta = 82.00$ s, 95% CI [-5.61 s, 169.61 s], $d = 1.28$, 95% CI [-0.22, 2.78], $p = 0.073$). Other post-hoc comparisons were not statistically significant. The most notable difference to Scenario 1 was that the *Baseline* variant was similarly fast as *Rec* and *Cont*. This effect can be explained by the difference in time criticality, as discussed in Section 5.1.2. In Scenario 1, pilots compared several options in greater detail, which was facilitated by AI support; whereas in Scenario 2, they tended to only check the nearest airport to speed up the decision. This short check could be performed quickly even without AI support.

There were no statistically significant differences in decision times in Scenario 3 ($F(3, 28) = 0.544, p = 0.66$), with an effect size close to zero. This again can be explained by the factor of time criticality. Participants had no time pressure in Scenario 3 and hence took their time for in-depth comparisons, irrespective of which system variant they used.

Overall, the most notable observation was that *Rec* was surprisingly not faster than *Cont*; it even tended to be slightly slower in the first two scenarios, though not significantly. Interestingly, the combination *Rec+Cont* did lead to a statistically significant speedup in Scenario 2. This is consistent with the usage patterns in Figure 9: Many participants in this group already prepared during normal flight for Casablanca in case of a time-critical emergency. When the emergency happened, the system confirmed their prepared plan (S10, forward reasoning), reducing the time to verify the plan. With *Rec*, participants got the same recommendation, but since they were less prepared, they had to verify the recommendation first (S3, S4, backward reasoning), which consumed a significant amount of time. The same trend can be observed in Scenario 1, though less pronounced, likely due to less time pressure and the disruption of handling the engine failure, as discussed in Section 5.1.2.

5.4 RQ4: Pilots' Perspectives

We identified three major themes in our thematic analysis of the exit interviews and occasional comments during the think-aloud sessions.

5.4.1 Greatest added value: making a lot of information quickly accessible. For participants, the greatest added value of the DAS was not the AI features per se, but rather the quick overview of a large amount of relevant information, as explained by 30 out of 32 participants. Today, pilots must gather information from various sources, which is time-consuming and error-prone. Our DAS improves on the current state even in the *Baseline* version by presenting all information at a glance in the table. Several participants mentioned that the table is basically what pilots in some airlines fill out by hand as part of FOR-DEC. Having it directly available speeds up decision-making and enables better decisions by reducing workload and providing more complete information:

> *"If you have time, you can work it all out in half an hour. But if you have a time-critical failure, [...] you need a lot of experience because you don't have the time to work it all out. You can spend hours doing FOR-DEC, but you only have fuel for half an hour. And then you have to make an intuitive decision based on your gut feeling. These are often not the best decisions."* (C2)

But AI features do contribute to the quick overview. While having all information at a glance is helpful, it is also overwhelming, as criticized by six participants. Pilots see the importance of AI in reducing clutter by surfacing relevant and hiding or de-emphasizing less relevant information. 14 pilots stressed that recommendations and ranking by AI evaluation help them to quickly focus on good options: *"I think this pre-filtering is important. And afterwards, the human can still fine-tune it a little."* (RC5). Also 14 participants explained how the color highlights help them to quickly identify potential limitations at airports: *"It's quite clever if you can see right away, do I even need to look at it or not?"* (RC2). Participants also suggested how AI could further reduce clutter, e.g. by hiding very bad options, or by evaluating weather as a whole and only displaying noteworthy weather aspects. In our system, all weather components are shown in the table and evaluated individually, creating a lot of redundancy: *"A little less information, visibility 10,000 or more, that's redundant information, that doesn't help me. So more with the question, what information creates added value?"* (R5).

Two participants cautioned against careless implementations of AI. They pointed out that especially the value of recommendations may be undermined by the need to double-check them, which could slow pilots down: *"You are inclined to question the computer, and you also want to decide for yourself, but at the same time you want to use everything, and then you'll probably need ages to figure it all out."* (B8). This view is consistent with the decision times in Section 5.3, where the *Rec* variant turned out slower than expected.

5.4.2 Tension between more system intelligence and more bias. While pilots welcomed AI support to surface relevant information, they were also concerned that too obtrusive AI could bias them toward subpar decisions:

> *"The more is presented to you and processed for you, the less you are actively involved. On the one hand, it helps a lot. On the other hand, you have to be careful whether all of this is exactly what you actually want."* (C5)

> *"It's a fine line as to whether you do too much."* (B1).

This tension was especially apparent with recommendations: 17 participants found recommendations helpful or suggested to add them if their system variant did not have them, while eleven participants rejected them or found them not helpful. Those who rejected recommendations felt that they remove pilots from the decision-making:

> *"I find it a bit difficult that the decision is given to you immediately like this [snaps fingers], and then you immediately go, ah okay, if it says so, then we'll take Hanover."* (RC4).

> *"What I would not like is for the AI to simply say 'Fly to Berlin!' or something, because then you don't know exactly where it's coming from. The human is the master in the system, the AI has to be subordinate and perform supporting tasks."* (B5)

Other participants saw no threat in recommendations to their agency, emphasizing that it is their job to monitor and question the system.

Hints and highlights were seen as less problematic and accepted by all participants, as they were perceived as *"less patronizing"* (R5): *"I have the feeling that I'm using the system, not that the system is using me."* (C2). However, seven pilots stressed that the highlights must make sense to be useful. Pilots especially disagreed with the red highlights of the "medium to good" braking action in Scenario 3, as described in Section 5.2.

Participants discussed several ways how AI may be introduced without removing pilots from the decision-making. For one, nine pilots noted that continuous support helps them to familiarize with the system and improve their SA, which also makes using the system in an emergency easier:

> *"It's important to be familiar with the system, because if you only use it in an emergency, it's like, oh yeah, is that right? If you're always plotting where you are, you already know a bit of what it says and you don't see, ah okay, it's red, it doesn't work."* (C8)

Moreover, participants had several ideas for alternative forms of AI support. Some suggested that the system only marks whether airports are suitable for a diversion or not, instead of making specific recommendations. Another idea was to display recommendations only after pilots have independently reached a decision, which some pilots likened to their collaborative decision-making:

> *"It's like with our CRM[14], when we work together, I have to be careful not to say as captain 'I think that airport is great, that's where we're flying to now, or would you mind?' Then of course he [the first officer] says 'No, let's do it.' That's why I always have to keep it open and not say what I've been thinking. Let the other one say it so that there is redundancy, or perhaps have my own mistakes pointed out."* (C5)

Another suggestion by several participants was that pilots have to manually define their criteria before they get recommendations. The system could also step in with suggestions for additional criteria that the pilot may have forgotten. RC2 proposed that AI could be used to evaluate the airports according to high-level categories like weather or operations. Pilots could then sort the airports by these high-level categories, rather than by low-level criteria like in our system. C8 advocated that the system should be limited to hard facts and leave soft factors to pilots:

> *"The wind won't change, that's for sure. The fuel won't change either. But I still have to see where the company wants us to go. Where do we perhaps know our way around? [...] So these soft factors, if they were included, I think that would be a bit like taking the decision away."* (C8)

However, opinions differed on this last suggestion, as B2 argued for the opposite and said that the system should be able to judge when a situation is so critical that it is acceptable to land somewhere *"even if I am now busting a limit."* (B2). The system should be able to *"classify, how serious is the incident for the danger of all people on board or for the aircraft?"* (B2).

While the discussions were predominantly about the risk of system-induced biases, four pilots brought the complementary perspective into the conversation, pointing out that AI could also help mitigate human errors. AI could for instance highlight good but less familiar options, and does not overlook things under time pressure.

5.4.3 Transparency and control: need for appropriation. Lastly, pilots discussed several ways in which they require transparency and control. Some participants wondered about how the AI works, e.g., why a certain recommendation is given (three participants), how the criteria are weighted (four participants), or why certain information is highlighted (four participants). However, the much more prevalent issue was what exactly the system considers for its evaluations (17 participants): Does *time to destination* include the time required for the approach? Does *fuel at destination* include final reserves? Does *stop margin* consider performance limitations after a technical failure? The predominant question was therefore not "How does the AI work?", as assumed by most research in explainable AI, but rather "How does the information fit my intention?".

Pilots further expressed the requirement to be able to control the AI. Five participants explicitly valued the option to edit the recommendation criteria to engage with the AI in negotiation patterns, as described in Section 5.1.1. Moreover, participants made further design suggestions to enable more control, such as options to pin airports to the top of the table and to manually hide airports, or to search for airports that are not in the table. R3 further suggested to add an extra column that pilots can fill themselves, for edge cases that are not covered by the system criteria.

[14] *Crew Resource Management*, procedures for effective communication and decision-making to prevent human error.

The common theme behind both transparency and control was that pilots want to be able to appropriate the system to fit their intention and momentary information needs. Most of the transparency and control requests were comments and questions interposed during the think-aloud sessions, where participants had an intention while using the system and asked for clarification to understand if the system would fit their intention.

6 Discussion

We discuss the takeaways of our results in Section 6.1 and Section 6.2 and bring them together in Section 6.3. We close with limitations and future work in Section 6.4.

6.1 There Is More to Decision Support than Giving Recommendations and Explanations

The common, but often tacit design goal for AI decision support is to solve the task for users, creating a redundant rather than complementary role to the human. While it is popular to speak of "human-AI collaboration" in AI-assisted decision-making research [29, 56], users may perceive this recommendation-centric support more as "human-AI competition" [73]. Some of our participants expressed similar concerns regarding recommendations.

Like previous studies involving real-world tasks with experts, our participants' views suggest a shift in the design goal from solving the entire decision task for users to addressing their primary pain points. For the tumor assessment use case studied by Lindvall et al. [42], pathologists' biggest challenge was to find small tumorous regions in huge images. In the case of sepsis diagnosis studied by Zhang et al. [73], physicians wanted to know which lab tests to order, when. In our case of diversions, pilots would benefit the most from a quick overview of a large amount of relevant information. This shift in design goal enables much more diverse uses for AI than merely giving end-to-end recommendations, as exemplified by the various suggestions by our participants.

The challenge shifts from giving best possible recommendations and explaining them, to serving pilots' momentary information needs. Transparency is consequently not only—or maybe not even primarily—required to calibrate users' reliance on the recommendation, as is the current focus in AI-assisted decision-making research [22]. Rather, the role of transparency is to make visible how well the information fits pilots' current intention.

Fine-grained control is further necessary to allow pilots to effectively cater the information to their intentions when the system-inferred information does not perfectly fit. This prevents the all-or-nothing situation often created by typical recommendation-centric support: Due to the closed nature of end-to-end recommendations, decision makers usually only have the choice to fully accept or reject the recommendation[15], which is also how users' reliance behavior is often modeled [54]. However, as observed by Sivaraman et al. with clinicians [59], decision makers' reliance behavior can be much more nuanced when they are given the opportunity. In their case, the AI recommendation consisted of several aspects, allowing clinicians to adopt some aspect while overruling another. With our DAS, pilots could go one step further by editing the recommendation criteria. This led to some productive uses of the system in instances where the pilot would have simply ignored or rejected the system had the controls not been available.

6.2 Recommendations Have to Be Embedded into Forward Support to Be Beneficial

According to our results, recommendation-centric support significantly constrains pilots from thinking beyond the limits of the system, which in line with previous work [6, 32, 60] leads to more overreliance. Recommendations on their own also surprisingly did not lead to faster decisions

[15]At least in classification tasks. In regression tasks, users can give the AI recommendation a more continuous weight in their decision-making [24, 47].

than continuous support, as the need to review the recommendations canceled out the gains from surfacing good options. Lastly, the use of recommendations was highly disputed among participants.

But while our results suggest that recommendations may not be the most valuable use of AI for diversion assistance, we did find benefits of providing recommendations. First, recommendations can serve as confirmation, which can accelerate decisions in time-critical situations, which was particularly evident in combination with continuous support in Scenario 2. Recommendations can further benefit decision-making when they challenge pilots' own ideas and trigger further considerations. This was more apparent in the interview statements, but could also be observed in one instance during the think-aloud sessions, namely with the participant that tried without success to align the recommendations with his own idea, as described in Section 5.1.1.

Note that both beneficial uses of recommendations require forward reasoning, as pilots must independently come up with a favored option to be confirmed or challenged. We found continuous support to be effective for promoting forward reasoning, but it lacked robustness against disruptions between normal flight and abnormal situations. Following such disruptions, pilots tended to revert to backward reasoning. To ensure reliable forward reasoning when presenting recommendations, additional measures are necessary, with potential suggestions provided by our participants.

6.3 From Recommendation-Centric to Process-Oriented Support

Bringing together the discussed aspects, we propose *process-oriented decision support* as a promising alternative framework to design AI decision support that is not centered around end-to-end recommendations. Figure 12 shows how process-oriented support compares to recommendation-centric support, and how it applies to the diversion use case through both continuous support and participants' design suggestions. We consider the framework to be applicable not only for diversion assistance but more generally for AI support of high-stakes decisions. We argue that process-oriented support is particularly useful for complex decisions where human expertise and context knowledge is crucial, but hard to combine with end-to-end recommendations, such as in healthcare [73], social work [35], or sales [4]. The key difference to recommendation-centric support is the shift from trying to solve the task *for* users through end-to-end recommendations, to *helping* users to solve the task by addressing the challenges in their decision-making process. These challenges have to be determined for each application through user research. In our diversion use case, pilots' main challenge was to gather the information they need to make a decision.

This shift in design goal has several implications. First, recommendation-centric support removes users from the decision-making process. It pushes users into backward reasoning, with the recommended end result as the starting point, leading to the lack of cognitive engagement and hence inappropriate reliance observed in prior work [8, 23]. This is not only an issue for pilots as in our study, but also in other domains, such as in healthcare [21, 32, 65]. Process-oriented support in contrast keeps users engaged in their decision-making process and helps them to generate a decision themselves while reasoning forward. Recommendations are optional, serving as confirmation or challenge toward the end of the process rather than the starting point. As reflected by our participants' discussions, the core challenge for process-oriented support is to strike a good balance between interpreting information with AI and leaving the interpretation work to the human, which is strongly use-case-dependent. Our continuous support concept leaves much of the interpretation to pilots, which worked well for the diversion use case since pilots are highly trained expert users who usually know what information is important, as demonstrated by participants in the Baseline condition. For such expert users, a restrained form of AI support like continuous support has the benefit that it allows them to work very flexibly with the system, as discussed by our participants (Section 5.4.2). Use cases with less expert users may require a "stronger" role for the AI, e.g., guiding

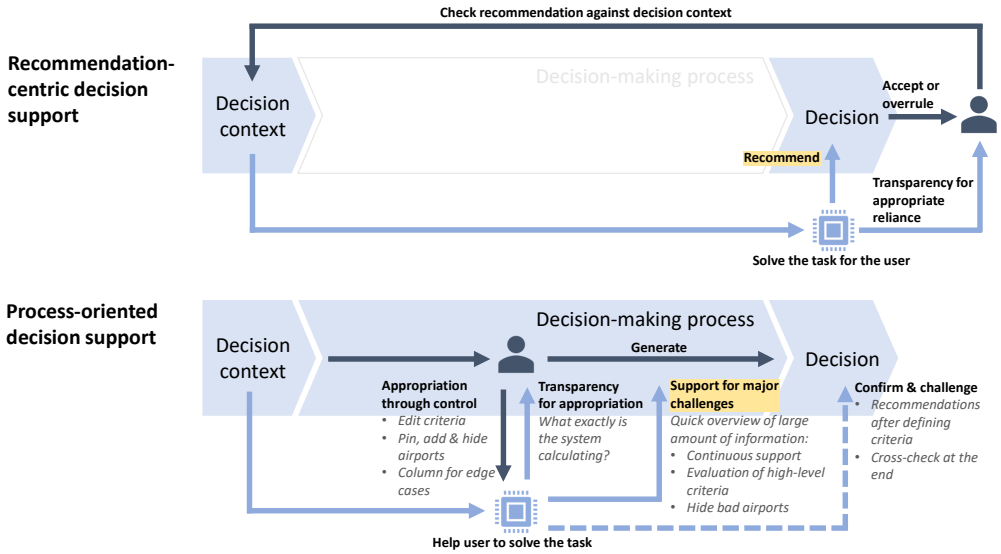


Fig. 12. Comparison of recommendation-centric and process-oriented decision support. The yellow highlights signal the core functionality of the AI in both paradigms. The dashed arrow indicates that decision recommendations are optional in process-oriented support. The italic text lays out how process-oriented support can be implemented concretely for the diversion use case, based on our results and participants' suggestions.

users toward important information through explanations. Even for our use case, some pilots wished for more sophisticated AI functionality.

Second, the role of transparency changes. In recommendation-centric support, transparency aids appropriate reliance on recommendations. In process-oriented support, transparency facilitates appropriation [15] by helping users understand how well the system functionality aligns with their momentary intentions and needs. The importance of appropriation in AI-assisted decision-making has been discussed by several authors recently [9, 18, 76]. It differs from reliance calibration in at least two aspects. One is that transparency for appropriation is targeting the more granular level of intermediary decision-making steps, rather than the end of the process. This leads to more nuanced behavior than a simple accept-reject dichotomy, like partial reliance and negotiation patterns, as observed by Sivaraman et al. [59] and with our participants as well. The other difference is that appropriate reliance concerns the correctness of the AI, while appropriation is about the user's intention. As was evident from our participants' requests, this does not necessarily involve explaining how the algorithm works.

Lastly, while control has been studied in AI systems for instance in the context of interactive machine learning [16, 37, 61], there is usually no control in recommendation-centric decision support. In process-oriented support, control is an important complement to transparency to enable appropriation. Notably, different to interactive machine learning, control is not necessarily about feedback that the model should learn, but rather about steering the system according to users' momentary intentions and context, as exemplified by the criteria editing feature in our system, or the controls designed by Cai et al. in a healthcare application [10].

Process-oriented support joins the ranks of recent alternatives to recommendation-centered support. In particular, we see it as a generalization of evaluative AI [45]. Core to process-oriented support is to identify the main challenges in a decision-making process which can be supported

by AI. Evaluative AI can be interpreted as process-oriented support for decisions where the main challenge is to generate and evaluate multiple hypotheses, such as in medical diagnostics. In the diversion use case, the main challenge was different, namely the difficulty of information gathering. This leads to different support opportunities, such as our continuous support concept, which proved to be a good, but not perfect starting point. Further research is required to better understand how each element of process-oriented support can be implemented for diversion assistance and beyond, from useful AI support roles, over how to enable appropriation through transparency and control, to how recommendations can be integrated toward the end of users' decision-making process.

6.4 Limitations and Future Work

There are several limitations to our results, mostly stemming from our flight simulation setup. For one, we decided to conduct the study with a single pilot at a time, rather than with two pilots as is the standard in today's cockpits. We are fully aware that this is a significant deviation from how pilots work today, as the cooperation of both pilots is fundamental to today's operations [28]. We still decided for a single-pilot setup for two reasons. First, recruiting pilots is challenging, and relying on the simultaneous availability of pairs of pilots would have significantly reduced our sample size or even made the study impossible to conduct. Second, we are interested in how different AI support paradigms would affect pilots' decision-making in a possible AI-augmented future of aviation. We are agnostic as to whether this future maintains a two-pilots cockpit or involves single-pilot operations, which is a long-standing goal of the aviation industry [26].

Another limitation was that we could not provide all tools that pilots use in their daily work, such as maps or tools to calculate certain performance and flight data. Besides being prohibitively costly to replicate all these tools, they are also not standardized between airlines. Still, participants were able to appropriate what was available to them in X-Plane to fulfill many of their information needs. Several participants also praised the study setup to be well executed. We therefore assume that the impact of the limited tools was not critical.

Furthermore, despite our best efforts to construct a representative range of valid scenarios, it was not possible in our setup to fully reflect complexities that can arise from the coordination with multiple stakeholders. Such coordination was rudimentarily present in Scenario 3, but in reality, these interactions would be much more open-ended and therefore hard to simulate in a controlled setting like ours. While studies like ours do produce valuable insights, other methods would be required to better capture the intricacies of such open-ended scenarios. For instance, one could analyze incident reports post deployment, which is a proven strategy in aviation to improve safety.

Lastly, given the difficulty and cost to recruit professional pilots, our sample size was relatively small for our quantitative analyses. We tried to address this limitation by triangulating the quantitative results with rich qualitative data from the think-aloud sessions and exit interviews. We found clear trends in the quantitative data that were consistent with the qualitative data.

Overall, these limitations highlight the difficulty of rigorously evaluating different support paradigms on a realistic task with domain experts. We suspect this is also the reason why such studies are rare. Nevertheless, similar to previous AI-assisted decision-making research on real-world tasks, our study produced insights that would likely not be available with simple tasks and crowd worker participants. We therefore encourage further research on realistic tasks with domain experts to close the gap between research and real-world adoption of AI decision support.

7 Conclusion

We conducted one of the first empirical comparisons between the dominant recommendation-centric AI decision support paradigm and alternative approaches in a realistic task setting with

domain experts. We found that with recommendation-only, participants exhibited strong overreliance as they thought beyond the limits of the system significantly less than with the baseline variant. The recommendation-only variant was also surprisingly not faster than continuous support, in fact it even tended to be slower. We further found that to benefit from recommendations, pilots must engage with them while reasoning forward. Continuous support appeared to be an effective approach to encourage forward reasoning, allowing to display recommendations with less overreliance and leading to faster decisions when combined with recommendations. However, the effectiveness of continuous support was sensitive to disruptions between normal flight and abnormal situations. Our qualitative analysis reveals that pilots do not primarily value the system for suggesting suitable options, but rather for helping them to gain a quick overview of a large amount of information. While pilots welcome AI features for this purpose, they are concerned that too obtrusive AI takes away too much of the decision from them, with recommendations being particularly controversial. Participants' statements further revealed their requirements for transparency and control to appropriate the system. Notably, their primary concern with transparency was not how the AI works, but rather how it fits their intention.

Our results challenge the assumption that AI decision support should be recommendation-centric and highlight the importance of supporting the decision-making process in a forward direction. Our continuous support concept was a promising first step in this direction, while our findings suggest many opportunities for improvement and future work. Further research is especially required to understand how to more robustly support decisions in a forward direction. We encapsulated our findings in a framework for process-oriented support, and envision our work as a contribution toward a more holistic perspective on AI-assisted decision-making that looks beyond recommendations and explanations.

Acknowledgments

This work was supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the LuFo VI-1 program, project KIEZ4-0.

Many thanks to our participants, who were eager to discuss our research with us, were happy to provide us with valuable contacts and insights into their daily work, and spread the word about our study among their colleagues.

References

[1] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels Van Berkel. 2023. "If I had all the time in the world": ophthalmologists' perceptions of anchoring bias mitigation in clinical AI support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 16:1–16:14. https://doi.org/10.1145/3544548.3581513

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama, Japan, 81:1–81:16. https://doi.org/10.1145/3411764.3445717

[3] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Analysing Data. In *Qualitative HCI Research: Going Behind the Scenes*, Ann Blandford, Dominic Furniss, and Stephann Makri (Eds.). Springer International Publishing, Cham, 51–60. https://doi.org/10.1007/978-3-031-02217-3_5

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

[5] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. 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)*. ACM, Hamburg, Germany, 15:1–15:19. https://doi.org/10.1145/3544548.3581251

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. 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)*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[7] Zana Buçinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z. Gajos. 2022. Beyond end predictions: stop putting machine learning first and design human-centered AI for decision support. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–4.

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

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

[10] Carrie J. Cai, Martin C. Stumpe, Michael Terry, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, and Greg S. Corrado. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, Scotland, UK, 4:1–4:14. https://doi.org/10.1145/3290605.3300234

[11] Longbing Cao. 2022. AI in finance: challenges, techniques, and opportunities. *Comput. Surveys* 55, 3 (Feb. 2022), 64:1–64:38. https://doi.org/10.1145/3502289

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

[13] Lingwei Cheng and Alexandra Chouldechova. 2023. 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)*. ACM, Hamburg, Germany, 756:1–756:27. https://doi.org/10.1145/3544548.3581253

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

[15] Alan Dix. 2007. Designing for appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers (BCS-HCI '07, Vol. 2)*. BCS Learning & Development Ltd., Lancaster, UK, 27–30. https://doi.org/10.14236/ewic/HCI2007.53

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

[17] EASA. 2023. *Artificial Intelligence Roadmap 2.0: A human-centric approach to AI in aviation*. Technical Report. European Union Aviation Safety Agency (EASA). 36 pages. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20

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

[19] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, Glasgow, Scotland, UK, LBW0243:1–LBW0243:6. https://doi.org/10.1145/3290607.3312787

[20] Sebastian S. Feger, Felix Ehrentraut, Christopher Katins, Philippe Palanque, and Thomas Kosch. 2022. HCI for general aviation: current state and research challenges. *Interactions* 29, 6 (Nov. 2022), 60–65. https://doi.org/10.1145/3564040

[21] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. 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)*. Association for Computing Machinery, Seoul, Republic of Korea, 1362–1374. https://doi.org/10.1145/3531146.3533193

[22] Raymond Fok and Daniel S. Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* Early View (July 2024), 1–16. https://doi.org/10.1002/aaai.12182

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

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

[25] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. 2023. Augmenting pathologists with NaviPath: design and evaluation of a human-AI collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 349:1–349:19. https://doi.org/10.1145/3544548.3580694

[26] Don Harris. 2007. A human-centred design agenda for the development of single crew operated commercial aircraft. *Aircraft Engineering and Aerospace Technology* 79, 5 (Sept. 2007), 518–526. https://doi.org/10.1108/00022660710780650

[27] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. 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 (Oct. 2023), 276:1–276:29. https://doi.org/10.1145/3610067

[28] Robert L. Helmreich, Ashleigh C. Merritt, and John A. Wilhelm. 1999. The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology* 9, 1 (1999), 19–32. https://doi.org/10.1207/s15327108ijap0901_2

[29] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, Sydney, NSW, Australia, 453–463. https://doi.org/10.1145/3581641.3584052

[30] Hans-Jürgen Hörmann. 1994. FOR-DEC - A prescriptive model for aeronautical decision making. In *Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP)*. Avebury Aviation, Dublin, Ireland, 17–23.

[31] 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. 2021. 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)*. ACM, Yokohama, Japan, 659:1–659:14. https://doi.org/10.1145/3411764.3445385

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

[33] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P. Wallach. 2020. "You have to piece the puzzle together": implications for designing decision support in intensive care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. ACM, Eindhoven, Netherlands, 1509–1522. https://doi.org/10.1145/3357236.3395436

[34] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. 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)*. ACM, New Orleans, LA, USA, 52:1–52:18. https://doi.org/10.1145/3491102.3517439

[35] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "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)*. ACM, Virtual Event, Australia, 454–470. https://doi.org/10.1145/3532106.3533556

[36] Sean Koon. 2022. A human-capabilities orientation for human-AI interaction design. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–5.

[37] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. Association for Computing Machinery, Atlanta, GA, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[38] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, Chicago, IL, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[39] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, Atlanta, GA, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[40] Bridget A. Lewis, Valerie J. Gawron, Ehsan Esmaeilzadeh, Ralf H. Mayer, Felipe Moreno-Hines, Neil Nerwich, and Paulo M. Alves. 2021. Data-driven estimation of the impact of diversions due to in-flight medical emergencies on flight delay and aircraft operating costs. *Aerospace Medicine and Human Performance* 92, 2 (Feb. 2021), 99–105. https://doi.org/10.3357/AMHP.5720.2021

[41] Brian Y. Lim, Joseph P. Cahaly, Chester Y. F. Sng, and Adam Chew. 2023. Diagrammatization: Rationalizing with diagrammatic AI explanations for abductive-deductive reasoning on hypotheses. https://doi.org/10.48550/arXiv.2302.01241 .

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

[43] Han Liu, Vivian Lai, and Chenhao Tan. 2021. 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 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[44] John M. McGuirl and Nadine B. Sarter. 2006. 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 (Dec. 2006), 656–665. https://doi.org/10.1518/001872006779166334

[45] Tim Miller. 2023. 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)*. ACM, Chicago, IL, USA, 333–342. https://doi.org/10.1145/3593013.3594001

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

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

[48] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. 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)*. ACM, Sydney, NSW, Australia, 379–396. https://doi.org/10.1145/3581641.3584033

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

[50] Stephan Raaijmakers. 2019. Artificial intelligence for law enforcement: challenges and opportunities. *IEEE Security & Privacy* 17, 5 (Sept. 2019), 74–77. https://doi.org/10.1109/MSEC.2019.2925649

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

[52] Amy Rechkemmer and Ming Yin. 2022. 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)*. Association for Computing Machinery, New Orleans, LA, USA, 535:1–535:14. https://doi.org/10.1145/3491102.3501967

[53] Emilie Roth, Devorah Klein, Christen Sushereba, Katie Ernst, and Lauren Militello. 2022. *Methods and measures to evaluate technologies that influence aviator decision making and situation awareness.* Contract Report USAARL-TECH-CR–2022-22. Applied Decision Science, Cincinnati, OH, USA. 80 pages.

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

[55] Stephan Schlögl, Claudia Postulka, Reinhard Bernsteiner, and Christian Ploder. 2019. Artificial intelligence tool penetration in business: adoption, challenges and fears. In *Knowledge Management in Organizations (KMO 2019)*. Springer International Publishing, Zamora, Spain, 259–270. https://doi.org/10.1007/978-3-030-21451-7_22

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

[57] Scott Shappell, Cristy Detwiler, Kali Holcomb, Carla Hackworth, Albert Boquet, and Douglas A. Wiegmann. 2007. Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 2 (April 2007), 227–242. https://doi.org/10.1518/001872007X312469 .

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

[59] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 754:1–754:18. https://doi.org/10.1145/3544548.3581075

[60] P.J. Smith, C.E. McCoy, and C. Layton. 1997. Brittleness in the design of cooperative problem-solving systems: the effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 27, 3 (May 1997), 360–371. https://doi.org/10.1109/3468.568744

[61] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No explainability without accountability: an empirical study of explanations and feedback in interactive

ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 497:1–497:13. https://doi.org/10.1145/3313831.3376624

[62] Niels van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2021. Designing visual markers for continuous artificial intelligence support: a colonoscopy case study. *ACM Transactions on Computing for Healthcare* 2, 1 (Dec. 2021), 7:1–7:24. https://doi.org/10.1145/3422156

[63] Julian Varghese. 2020. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visceral Medicine* 36, 6 (Oct. 2020), 443–449. https://doi.org/10.1159/000511930

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

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

[66] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[67] David D. Woods. 1986. Paradigms for intelligent decision support. In *Intelligent Decision Support in Process Environments (NATO ASI Series, Vol. 21)*. Springer Berlin, Heidelberg, San Miniato, Italy, 153–173. https://doi.org/10.1007/978-3-642-50329-0_11

[68] Jakob Würfel, Boris Djartov, Anne Papenfuß, and Matthias Wies. 2023. 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 (AHFE 2023, Vol. 95)*. AHFE Open Acces, San Francisco, CA, USA, 589–597. https://doi.org/10.54941/ahfe1003844

[69] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. 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)*. ACM, Hamburg, Germany, 14:1–14:14. https://doi.org/10.1145/3544548.3581393

[70] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. 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)*. ACM, Glasgow, Scotland, UK, 238:1–238:11. https://doi.org/10.1145/3290605.3300468

[71] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F. Antaki. 2016. 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)*. ACM, San Jose, CA, USA, 4477–4488. https://doi.org/10.1145/2858036.2858373

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

[73] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. 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)*. ACM, Honolulu, HI, USA, 445:1–445:18. https://doi.org/10.1145/3613904.3642343

[74] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM, Barcelona, Spain, 295–305. https://doi.org/10.1145/3351095.3372852

[75] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. 2021. 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 (CHItaly '21)*. ACM, Bolzano, Italy, 18:1–18:5. https://doi.org/10.1145/3464385.3464696

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

A Exit Interview Guide

- How was your impression of the system?
- On a scale from 1—*not helpful at all* to 5—*extremely helpful*, how do you rate the helpfulness of the system?
- What is the greatest added value of the system?

- What do you find problematic or in need of improvement about the system?
- What was your strategy for using the system? What did you use it for?
- *For Rec, Cont, Rec+Cont conditions*: How important was/were the recommendations/normal flight mode/color highlights for your usage?
- *For Baseline condition*: What kind of AI features would you like to see in the system?
- Anything else that you want to ask or comment on?

B Participant Details

Table 3. Focus group participant details.

ID	Gender	Rank	Flight hours	Participated in pilot test
F1	Male	F/O	400	
F2	Male	Captain	10 200	
F3	Male	F/O	1600	X
F4	Male	F/O	370	X
Median			1000	
(IQR)			(392.5–3750)	

Received January 2024; revised April 2024; accepted May 2024

Table 4. Study participant details.

ID	Gender	Rank	Flight hours	# Diversions
B1	Male	F/O	3180	0
B2	Male	F/O	4000	4
B3	Male	F/O	2000	4
B4	Male	F/O	2000	1
B5	Male	F/O	200	0
B6	Male	F/O	1700	2
B7	Male	F/O	4800	3
B8	Male	F/O	5000	1
R1	Male	F/O	4000	2
R2	Male	F/O	2000	1
R3	Male	F/O	2500	1
R4	Male	F/O	4000	5
R5	Male	F/O	3000	0
R6	Male	F/O	1640	0
R7	Male	Captain	11 000	1
R8	Male	F/O	2500	1
C1	Male	F/O	2400	0
C2	Male	F/O	3000	0
C3	Male	F/O	6000	7
C4	Male	F/O	2500	2
C5	Male	Captain	13 000	5
C6	Male	F/O	2500	0
C7	Male	F/O	4000	5
C8	Female	F/O	580	0
RC1	Male	F/O	3800	3
RC2	Male	F/O	4000	0
RC3	Male	F/O	2000	3
RC4	Male	F/O	2400	2
RC5	Male	F/O	1900	3
RC6	Male	F/O	800	0
RC7	Female	F/O	2500	2
RC8	Male	F/O	3000	2
Median (IQR)			2500 (2000–4000)	1.5 (0–3)