Resilience Through Appropriation: Pilots' View on Complex Decision Support

Zelun Tony Zhang*
zhang@fortiss.org
fortiss GmbH, Research Institute of the Free State of Bavaria
Munich, Germany

Cara Storath*†
cara.storath@argodesign.com
argodesign
Munich, Germany

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

ABSTRACT

Intelligent decision support tools (DSTs) hold the promise to improve the quality of human decision-making in challenging situations like diversions in aviation. To achieve these improvements, a common goal in DST design is to calibrate decision makers' trust in the system. However, this perspective is mostly informed by controlled studies and might not fully reflect the real-world complexity of diversions. In order to understand how DSTs can be beneficial in the view of those who have the best understanding of the complexity of diversions, we interviewed professional pilots. To facilitate discussions, we built two low-fidelity prototypes, each representing a different role a DST could assume: (a) actively suggesting and ranking airports based on pilot-specified criteria, and (b) unobtrusively hinting at data points the pilot should be aware of. We find that while pilots would not blindly trust a DST, they at the same time reject deliberate trust calibration in the moment of the decision. We revisit appropriation as a lens to understand this seeming contradiction as well as a range of means to enable appropriation. Aside from the commonly considered need for transparency, these include directability and continuous support throughout the entire decision process. Based on our design exploration, we encourage to expand the view on DST design beyond trust calibration at the point of the actual decision.

CCS CONCEPTS

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

*Both authors contributed equally to this research.
†This work was completed while the second author was working at fortiss GmbH.

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03...$15.00
https://doi.org/10.1145/3581641.3584056

KEYWORDS

human-AI interaction, decision support tools, intelligent decision support, AI-assisted decision-making, naturalistic decision-making, imperfect AI, appropriation, aviation

ACM Reference Format:
Zelun Tony Zhang, Cara Storath, Yuanting Liu, and Andreas Butz. 2023. Resilience Through Appropriation: Pilots' View on Complex Decision Support. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3581641.3584056

1 INTRODUCTION

As with most industries, the aviation industry is currently pushing for the adoption of *artificial intelligence* (AI) [16, 29]. A generally popular application for AI are intelligent *decision support tools* (DSTs) [41], with the promise of improving the quality of human decision-making. Such improvements could be of great benefit to commercial aviation as well, where faulty decision-making is among the leading causes for accidents [52]. One particular type of decision situations for pilots are *diversions*, where an emergency or other abnormal event requires the crew to divert to another destination than the original one. While diversions are rare, they can be very challenging. For instance, diversions can incur high cost on airlines and strongly disrupt operations [42]; but minimizing these costs can run counter to safety requirements. A well-known example of such a goal conflict leading to a poor decision is Hapag-Lloyd Flight 3378 from Crete, Greece to Hanover, Germany [47]: After departure, the landing gear did not fully retract, causing increased fuel consumption. Instead of landing at nearby Zagreb, the captain tried to reach Vienna, likely because it was economically and operationally preferable. Unfortunately, the fuel did not last, resulting in a crash landing before the runway at Vienna. Supporting pilots to minimize the risk of such poor diversion decisions is a desirable goal given their potentially grave consequences. But in contrast to often simple controlled studies on DSTs, which treat decisions as isolated selections between options [41], diversions are highly complex, posing a challenge for the application of DSTs.

Currently existing AI applications in aviation are not safety-critical [29]. To move into safety-critical applications like diversion assistance, the question is how to keep the very high safety standards of the industry with AI, or even improve on them. While the introduction of flight deck automation has brought many benefits, it has also led to a range of unanticipated issues like e.g. *mode confusion* (user errors due to automation being in a different mode than expected) or *clumsy automation* (automation that eases workload when it is already low, but further increases workload when it is already high) [3, 22, 50]. The concern is that increased system intelligence might aggravate problems like these [19].

Such problems can often be attributed to an insufficient consideration of user requirements in the development of automation, despite the popularity of the label "human-centered" [27, 50]. As target applications for intelligent systems are often highly complex, a deep understanding of user requirements becomes even more important than with traditional automation. We therefore interviewed professional pilots to learn about the context of diversion decisions as well as opportunities and requirements for DSTs in diversions. To facilitate discussions, we built two low-fidelity prototypes, each representing a different role a DST could assume: (a) actively suggesting and ranking airports based on criteria the pilot has previously defined, and (b) unobtrusively hinting at data points the pilot should be aware of.

We find that diversion decision-making goes beyond what is usually captured in controlled studies. Most importantly, pilots do not merely try to select a good option in reaction to an emergency, but also aim to proactively shape the situation in their favor. We also find that while pilots would not blindly trust a DST, they at the same time reject deliberate *trust calibration* in the moment of the decision. We revisit the concept of *appropriation* [14] to make sense of this seeming contradiction. Together with the insights into diversions, the lens of appropriation suggests a range of design opportunities and challenges which in part go beyond the common perspective on DSTs (i.e. recommending a decision and ensuring trust calibration through transparency). Most notably, this includes continuous support prior to and around the actual point of decision. We view the contributions of our work as follows:

(1) We contribute to and demonstrate the value of the growing efforts to investigate intelligent DSTs in complex real-world applications, with aviation as a domain that is understudied in the HCI community.
(2) We propose to design DSTs and intelligent systems more generally with appropriation in mind, and discuss possible means to facilitate appropriation of intelligent systems.

2 BACKGROUND

2.1 Human decision-making

Early research on human decision-making was dominated by *rational choice theory*, i.e. the assumption that humans would in general make rational decisions given the available information [23]. Later evidence undermined this assumption, showing that humans frequently deviate from "optimal" decisions predicted by formal models [23, 24, 43]. As a result, rational choice theory has given way to the *dual process theory of decision-making* [34]. According to this model, humans employ either intuitive, fast decision-making (*System 1*), or deliberative, slow decision-making (*System 2*). The mechanisms underlying the intuitive System 1 judgments and their interactions with System 2 have been of great interest to decision-making research, with at least two established perspectives: *heuristics and biases* (HB) and *naturalistic decision making* (NDM) [35]. The HB approach focuses on systematic biases and errors in human intuition, mostly investigated through controlled experiments [35, 56]. NDM on the other hand focuses on how experts are able to make good decisions intuitively, studying decision-making under real-world conditions and complexities [39, 43]. NDM views decisions much more broadly than only as the choice between options, with a focus on *macrocognitive functions* like *sensemaking* [49]. In fact, according to NDM models like *recognition-primed decision* (RPD) [38], expert decisions may not involve choosing between options at all. RPD—"the prototypical NDM model" [43]—posits that experts generate one option at a time that they recognize as appropriate to the current situation and that they evaluate through mental simulation. Only if an option is deemed to have shortcomings, the expert would discard it and generate a new one.

2.2 Decision-making in aviation

Decision-making in emergency or abnormal situations is one of the most challenging tasks for pilots, as a multitude of technical, operational, and environmental factors need to be considered. For instance, pilots may need to consider technical limitations of the aircraft due to system failures, preferences of the airline, and current weather conditions. Since goals can be conflicting, there is often no objectively correct decision. Moreover, decisions in aviation are often accompanied by time constraints and uncertainty, e.g. about how weather conditions will develop. To make good decisions in such complex and dynamic situations, it is crucial for pilots to form and maintain *situation awareness* (SA). According to the commonly adopted model of Endsley [18], SA consists of three levels: 1) *perception* of elements in the current situation, 2) *comprehension* of the current situation, and 3) *projection* of the future status.

Poor decisions like in the example of Hapag-Lloyd Flight 3378 are strongly linked with low SA on the pilots' side [18]. Often, this is a result of unstructured decision-making based on faulty intuition. In order to help pilots make structured decisions with a high level of SA, airlines have introduced multiple prescriptive decision frameworks [55]. While these frameworks differ in details, they have in common that they structure decisions along several steps, such as analysis of the situation or evaluation of options [55]. In Germany and other European countries, one of the most widely used frameworks is *FOR-DEC* [31, 55]. The acronym concisely captures the six steps that pilots should follow in their decision-making process: facts, options, risks and benefits, decision, execution, check. The dash in the middle symbolizes a pause after the situation assessment phase before going for a specific option.

In this work, we focus on diversions, where an emergency or abnormal situation requires the crew to divert to an airport different from the original destination. In such a case, pilots need to first decide whether a diversion is necessary, and if so, which alternative destination to divert to. Possible reasons for a diversion include e.g. bad weather at the destination, a technical failure, or a medical emergency among the passengers.

2.3 Decision support tools

Along with the impressive advances in AI, interest in intelligent *decision support tools* (DSTs) has surged in recent years, often in high-stakes applications like medical diagnosis [5, 33, 48], creditworthiness assessment [11, 25], university admission [9, 13], or recidivism prediction [25, 44, 58]. Usually, these DSTs feature an AI model that recommends a decision, often along with an explanation [41]. As Cabitza et al. [7] call it, the AI functions as an "oracle".

Reflecting the HB tradition of decision-making research, there has recently been a quickly growing number of controlled studies on intelligent DSTs, often through the lens of human cognitive biases [6, 11, 46]. A frequent focus of these studies is *trust calibration* via explanations of AI outputs [51, 58, 60, 63], i.e. whether and to what extent explanations help decision makers to adopt correct outputs and to reject incorrect ones. The reasoning is that well-calibrated trust would lead to high joint human-AI performance [2]. Results show that trust is hard to calibrate, which potentially results in *overreliance* [2, 5, 33] or *underreliance* [13].

While insightful, these studies may not fully reflect the usage of DSTs in practice. Studies on trust calibration mostly frame the human-DST interaction as what is essentially *supervisory control*: Subjects have to detect when to trust the DST and when not to. Chiou and Lee [10] argue that trust calibration is indeed crucial for supervisory control, but not necessarily the goal for the often more lateral interactions between humans and intelligent systems. These lateral interactions are usually not captured by controlled studies. Similarly, Lai et al. [41] point out that controlled studies usually miss context, mostly treating decisions as isolated points, which might be hard to translate into practice.

The HCI community has studied actual decision environments as well, though, like clinical decisions [32, 36, 62], child maltreatment screening [37], or nursing homes [26]. Challenges revealed in such real-world studies include for instance a mismatch between model prediction and decision makers' thought process [37], or inaccurate assumptions about how decision makers would involve DSTs into their work [62]. Challenges like these show that how to best support high-stakes decisions in practice remains an open question and can likely not be fully addressed by controlled studies only. Within the aviation context, past research on DSTs emphasizes the issue of *brittleness* [12, 45, 54], i.e. how intelligent systems might fail under complex conditions and how this impacts operators. Prior work also points out that with regard to AI on flight decks, pilots themselves are most concerned about how AI might fail when faced with real-world complexities [64]. For instance, rules and procedures help pilots to deal with most situations. However, unusual events can require pilots to deviate from rules and procedures, a difficult judgment for computers to make. With this research context in mind, we posed the following research question for our design exploration with pilots:

> **RQ**: What are design opportunities and requirements for intelligent systems to support pilots in making diversion decisions, especially considering the high complexity of diversions?

3 METHOD

To understand how to support pilots in diversions, we created two low-fidelity prototypes of an intelligent diversion assistance system and discussed them with professional pilots. Below, we present our prototypes and the design rationale behind them, followed by a description of the pilot interviews.

3.1 Scenario

The scenario presented to the pilots was a passenger having a heart attack. In such an emergency, pilots should land the aircraft soon to enable medical care. This most likely means diverting to another airport in case the planned destination is not close enough. The prototypes in theory should not only apply to diversions due to medical emergencies, but also other abnormal situations, such as technical failure. For this design exploration however, only the user journey of a medical emergency was laid out in detail.

3.2 Prototypes

We built two click dummy prototypes for diversion assistance systems. They mainly differ in the role of the intelligent system and how the pilot can interact with it. Note that both concepts are not mutually exclusive but could easily be combined. We chose to draft them as two separate concepts in order to emphasize during the discussions with pilots the different roles a DST could assume.

3.2.1 Design process. We gained an initial understanding about diversions through informal interviews with pilots and industry experts as well as through an existing diversion assistance concept from an industry partner [20]. Information gained from these sources include common decision criteria, pilots' general workflow during diversions, as well as rough ideas about the difficulties of diversions, such as the inefficiency of collecting relevant information for a decision.

Based on this initial understanding, we created the prototypes using a diverge-and-converge process [21]. The first two authors independently created several rough sketches of possible designs (diverge) before presenting them to each other and discussing them (converge). Prior work indicates that pilots are particularly concerned that an intelligent system might be a burden rather than a help in case it cannot adequately handle real-world complexities [64]. We therefore had a particular focus on how system intelligence can be helpful to pilots, even when it is imperfect at times. The major point of discussion that emerged from this diverge-and-converge exploration was how to combine or trade off between efficiency and control: The closer the DST output would be to a ready-to-adopt decision, the more efficient the support would be if the output was of high quality. At the same time, it would likely be harder for pilots to steer the DST into a more productive direction if the output was of low quality. The ideas from the sketches converged on two potential roles of a diversion DST, around which we built the two click dummy prototypes. The prototypes matured by reflecting on the design decisions taken and giving feedback to each other.

3.2.2 Prototype A—Global Suggestions. Prototype A is conceptually closer to the aforementioned industry partner design [20]. It follows the conventional approach for DSTs where the system gives

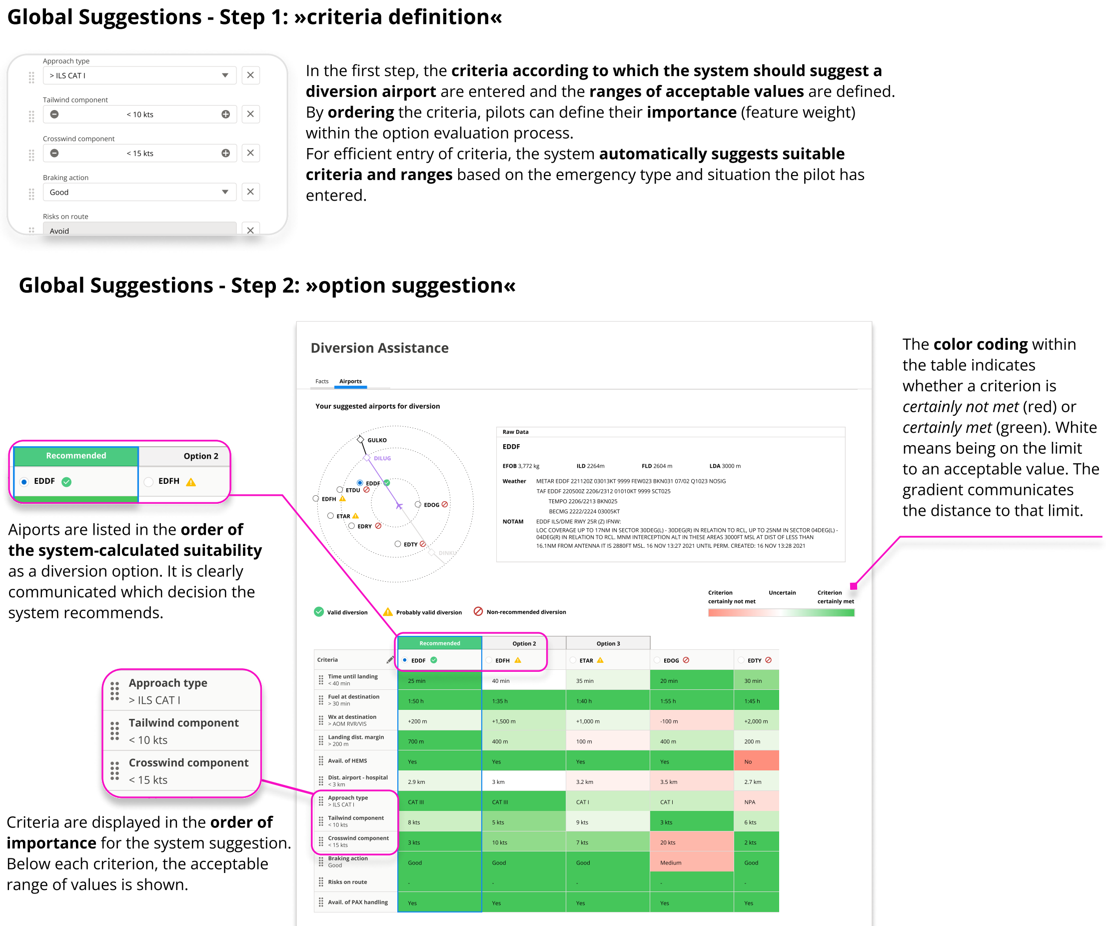

Figure 1: Prototype A—Global Suggestions (Section 3.2.2). Following the conventional DST approach, the system suggests airports by ranking them according to their suitability. Pilots have control over the ranking through the specification of diversion criteria. The color coding provides transparency.

concrete decision suggestions. In this case, the system gives suggestions by ranking the surrounding airports according to their calculated suitability and highlighting the best ones (Figure 1). Before getting a suggestion, pilots need to specify diversion criteria for the system to consider, such as flight time, runway length, weather, or distance to the next hospital. Pilots can further define the importance of the individual criteria and specify an acceptable range of values for them. Suitable criteria and values are suggested by the system based on the selected emergency situation (i.e. medical emergency, technical failure, etc.) in order to speed up the input process. Letting pilots define and check the criteria first before displaying a suggestion aims to provide more control over the system's evaluation. On the following screen, the ranked airports are presented in a table, along with their calculated values for the specified criteria. A color coding communicates how likely the corresponding criterion will be met for the respective airport. The values for the criteria together with the color coding are meant to provide transparency about the system's suggestions. Transparency together with control over the criteria are the two means in this prototype to help pilots work with possibly imperfect system suggestions.

3.2.3 Prototype B—Local Hints. In prototype B, the system assumes a different role than in conventional DST designs. Instead of suggesting diversion options, the system is designed to provide a better basis for a decision by continuously evaluating and highlighting relevant information (Figure 2). Similar to prototype A, this is done by displaying the surrounding airports and their information regarding various diversion criteria in a table. However, unlike in the other concept, prototype B does not rank the airports, and it does not require pilots to specify diversion criteria. Instead, the values for all criteria known to the system are shown at all times. For values that require extra attention, the system displays warnings and alerts called *Local Hints*. Clicking on a hint reveals an explanation of its reason. Pilots can additionally select an airport to see a summary box with a short communication of the main points highlighted by the system for the selected airport.

In contrast to prototype A, which is designed to be used when an emergency occurs, prototype B is meant to be used throughout

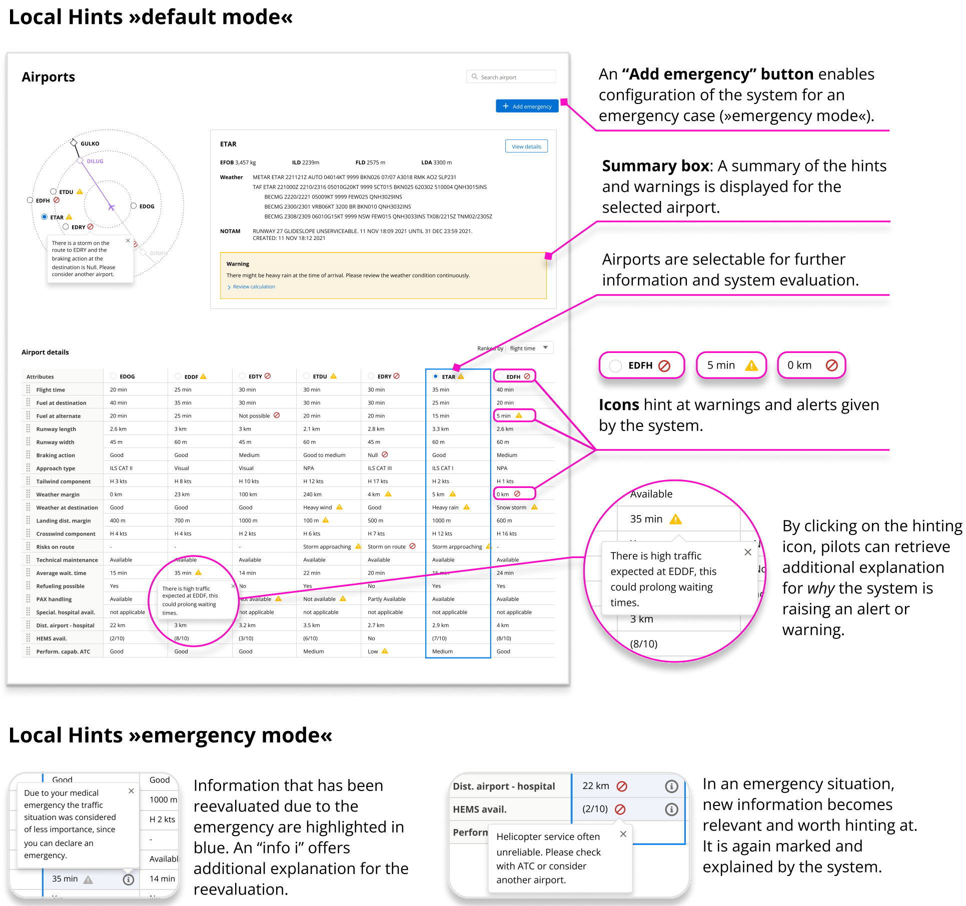

Figure 2: Prototype B—Local Hints (Section 3.2.3). Instead of suggesting airports, the system continuously alerts and warns about data points that require extra attention, also in normal flight. In case of an emergency, the hints can be tailored to the specific type of emergency. Text explanations provide transparency about the hints.

the flight, even when no emergency is imminent. It can be used in two modes: default mode and emergency mode. In default mode, the information is evaluated generally, with no particular reason for diversion in mind. This mode aims to provide SA, by giving pilots an overview of their surroundings and potential risks like bad weather which are of general interest, independent of the specific situation. In case an emergency occurs, pilots can activate the emergency mode and select a specific diversion reason, in this case medical emergency. The system then reevaluates all the data and gives hints tailored to the newly specified situation. For a medical emergency, this means for example displaying a warning if the helicopter service at an airport is predicted to be unavailable. In case the evaluation has changed for a certain criterion due to the emergency mode, the corresponding table cell is indicated visually.

In this prototype, we aimed to design around possibly imperfect system intelligence by giving the system a more restrained role. Our intention was to avoid biasing pilots toward possibly suboptimal options and to engage pilots more deeply into the solution generation. The idea is that since system intelligence is provided at a more fine-grained level, pilots might be able to better incorporate it into their decision-making, even when it is imperfect.

3.3 Expert interviews

We interviewed eight experienced pilots (seven male and one female, average age: 43.5 years, average flying hours: 10,279 hours). One of them has a background as a fighter pilot, but also regularly test-flies passenger aircraft. The other participants are airline pilots who have flown for various German airlines, five as captains, two as first officers. The airline pilots were recruited over mailing lists and received a compensation of 200 EUR each for their participation in the about two-hour-long sessions, which is a typical amount for pilots, given the difficulty of addressing this target group. The fighter pilot participated without compensation as an employee of a project partner. We conducted the interviews via Webex[1] video calls and recorded all sessions. Each session consisted of two parts, as shown in Figure 3 and as described below.

[1] https://www.webex.com

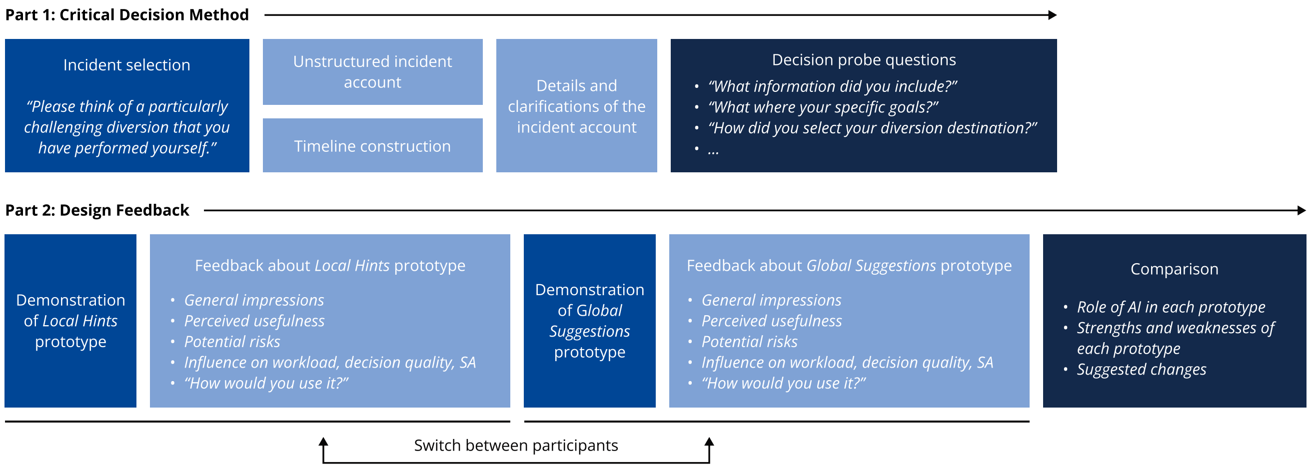

Figure 3: Overview of the interview sessions. Each session consisted of two parts: A short, half-hour interview about diversions according to the Critical Decision Method (Section 3.3.1), and a feedback discussion about our two prototypes (Section 3.3.2).

3.3.1 Critical Decision Method interview. We started each session with a half-hour semi-structured interview according to a pared-down version of the *Critical Decision Method* (CDM) [40], a method that is widely used especially in NDM research to elicit expert knowledge about complex decision-making tasks. We asked participants to think about one particularly challenging diversion from their own experience. If a pilot had no personal diversion experience, we asked them to think of a relevant simulator training instead. We then asked for a brief description of the incident, from the moment the pilot noticed the problem, until the completion of the diversion. While participants gave their account, we simultaneously took note in the form of a rough visual timeline. After they had finished their account, we showed this timeline to participants via screen sharing for them to clarify details or add missing pieces of information. Lastly, based on this account, we probed for more decision-making details using probing questions taken from [40].

This CDM-based interview served two purposes: For one, we wanted to gain a deeper understanding of the operational complexities of diversions. Additionally, their own specific experience served as a concrete example that participants could refer to and elaborate on during the second part of the session.

3.3.2 Feedback discussion. In the second part of the sessions, we discussed our two prototypes with participants, again in the form of semi-structured interviews. By confronting pilots with two different forms of decision support, we aimed to get feedback on the pros and cons of each, learn more about potential design opportunities and challenges, and about diversion decisions in general. We first discussed both prototypes separately before asking about them in comparison. Between participants, we switched the order of the prototypes to mitigate order effects. For each prototype, we first demonstrated it to participants through screen sharing by showcasing a typical user journey for our scenario. Afterward, we asked participants for their feedback. Besides questions about first impressions or useful features and potential risks, we also asked about the influence of such a system on workload, decision quality, and SA in situations of differing degrees of risk and urgency. One of our particular concerns was that DSTs might produce inadequate outputs, for instance due to events that the system is not designed to account for. As a further probe, we therefore first asked participants to describe how they would use the system for a diversion decision. Following that, we asked how their usage would look like in a situation where an important factor is not considered by the system. For this, we let them assume that the destination appearing most favorable based on the system outputs would be unavailable due to a no-fly zone caused by political disturbances. After discussing both variants in the described manner, we closed the interviews by asking participants how they perceived the role of the system in both, letting them compare the strengths and weaknesses of each, and asking for suggested changes to the prototypes.

4 RESULTS

We transcribed the recordings and analyzed the data through affinity diagramming [30], as our goal was to understand pilots' views on diversion assistance in a bottom-up manner. The first two authors split the recordings among them and independently transcribed one set of recordings each. While transcribing, both authors collected statements of interest as sticky notes on a Miro[2] board. The sticky notes were annotated with participant numbers and time stamps so that statements could be traced back to their source. During multiple affinity diagramming sessions, the first two authors clustered the sticky notes into low-level themes, initially grouping the themes under the CDM part, prototype A and B, and the comparison between both, according to the structure of the interviews described in Section 3.3. In a second step, we detached these low-level themes from the interview structure and clustered them further into higher-level themes that span across individual interview parts. Clustering decisions were made under constant discussion between the first two authors, until consensus was reached. The results are summarized in Figure 4. We ended up organizing them by pilots' goals and actions, their challenges, and means to support diversion decisions.

[2] https://miro.com

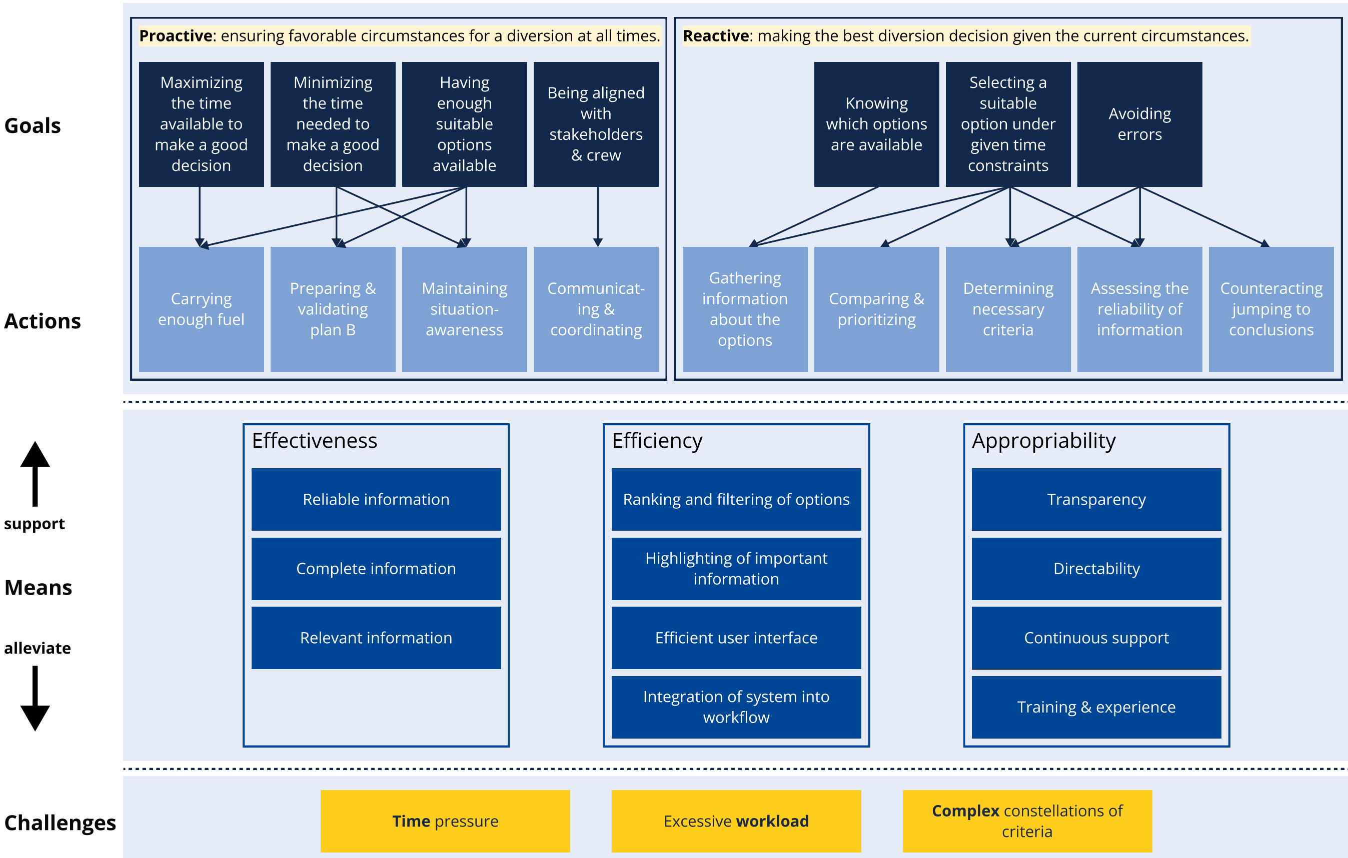

Figure 4: Overview of the interview results. Participants discussed their goals and actions (Section 4.1), challenges they face during diversions (Section 4.2), as well as possible means to support goals and actions and to alleviate challenges (Section 4.3).

4.1 Goals and actions

When an emergency or abnormal situation arises, pilots need to select an appropriate course of action in *reaction* to the situation. However, they do not only act reactively, but also aim to *proactively* shape the situation in their favor. This reflects the procedural nature of decision-making as emphasized by NDM (Section 2.1). Note that proactive and reactive goals and actions are not consecutive, but can (and often do) occur in parallel, as shown in Figure 5.

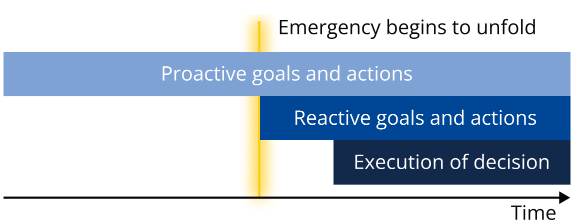

Figure 5: Temporal relationship between proactive and reactive actions and goals. The blur around the vertical line indicates that it is not always perfectly clear when an emergency begins.

4.1.1 Reactive goals and actions. The time constraints of the situation are the determining factor for how pilots select an option: In time-critical situations, they mostly only check if the first available option is feasible, with safety being the top priority, which conforms to the RPD model (Section 2.1). With more time, other factors like operational considerations can be considered, and multiple options can be compared in more detail.

Information gathering, or sensemaking in NDM terms, is an active process, in which pilots try to reduce uncertainty, although sometimes it is not possible to eliminate all uncertainty:

> *"Many views are unfortunately closed to us and we have to open the view ourselves quite laboriously."* (P2)
> *"If I had assumed that everything would go smoothly in Hanover and I would not expect long holding procedures[3] etc. and could land, then I could have flown there. But I didn't have the guarantee."* (P8)

Information about airports can be gathered through on-board systems or documentations, pre-flight briefing, and experience. Pilots are trained to avoid errors in their decision-making, e.g. by double-checking information, challenging each other, or avoiding jumping

[3] A procedure to temporarily keep aircraft in the air which have arrived at the destination but cannot land immediately, in this case due to high traffic.

to conclusions, as encouraged by frameworks like FOR-DEC (Section 2.2). All participants mentioned FOR-DEC, without us asking about it. Furthermore, decision-making is not done once an airport is chosen. Instead, pilots check for the validity of their choice as the situation evolves, represented as "C" in FOR-DEC.

4.1.2 Proactive goals and actions. As time is critical, pilots try to maximize the time available as well as to minimize the time needed for a good decision. This already starts before the flight. For instance, if it was foreseeable that the weather at the destination will be bad, they would carry extra fuel in order to have more reserves should landing at the destination not be possible. Pilots further always prepare an alternate to the destination as plan B, which can be quickly executed when needed. During the flight, they constantly have to make sure that their plan B is still valid and that they maintain SA of their surroundings: *"The challenge is always to maintain situational awareness. If I know where I am, I can be where I want."* (P3).

Strongly related to the time-related goals, pilots also aim to ensure that they always have enough suitable options available by anticipating possible developments (SA level 3). This means for instance in case of uncertainty about how a situation will evolve, pilots will choose a course of action where the risk of running out of options is low. For example, one pilot explained why he landed at an airport en route and not in Hanover, which was the originally planned alternate to the destination Hamburg:

> *"If half of the Hamburg traffic additionally diverts to Hanover, it can of course happen that some holding procedures are assigned, which would then force me into minimum fuel and to commit myself to staying in Hanover, with no choice left in my options."* (P8)

Lastly, diversions involve multiple stakeholders, including the two pilots, the cabin crew, air traffic control, and the airline. In order to facilitate good decisions, pilots have to actively communicate throughout the flight to align with all stakeholders.

4.2 Challenges

Good, structured decisions as prescribed by FOR-DEC take time. Under time pressure, the process has to be shortened, which might lead to worse decisions: *"The less time I have [...], the less options I have, and the worse the decisions get."* (P3). Time criticality is often determined by the amount of fuel left. Other time-critical situations include medical emergencies or technical failures.

Aggravating the issue of time pressure is the excessive workload in an emergency. For one, diversions involve many tasks besides the actual decision: *"I have to make clear to the controller what I am up to; I have to reselect an appropriate course; and so on!"* (P1). Moreover, cockpit systems are often cumbersome to use and not well integrated:

> *"We have these individual pieces of information, we have the slips of paper here that we then print out from our little computer—they're so tiny—for the weather [...]. Then I have the information that I get from the approach charts. [...] I have to call them up one by one, etc. etc. etc. [...] That's actually super time-consuming and you forget or you simply overlook things."* (P4)

However, modern aircraft have gotten much better in this regard, with information displays offering some degree of integration: *"Over the last ten years, there has been worlds of progress, worlds."* (P3).

Another challenge are the numerous criteria that might be important for diversions, possibly creating complex constellations. Some of the examples mentioned by participants include: weather (mentioned by all pilots) with multiple sub-criteria like wind or visibility; distance (P2, P4, P5, P7); fuel limits (P1, P3, P8); political situation (P3, P6); overland connection to original destination (P2, P8); number and size of runways (P1, P4, P6, P7); free parking spots (P1, P6, P8); aircraft-specific equipment (P2, P3, P6); company-specific infrastructure (P2, P6, P8); and passenger handling (P2, P8). However, the challenge is not only the number of criteria. Sometimes, these criteria also require very detailed and specific information:

> *"For example, landing gear does not lower. I would need a foam carpet. Can the fire department there do that? Or do I have to get the airplane onto the tarmac with the belly and the engines? [...] Does it even have a fire department that is capable enough to evacuate and extinguish an airplane with 180 people?"* (P8)

This complexity of possibly relevant factors makes it difficult for an intelligent system to cover every eventuality: *"Such things, they all have to be part of the decision-making process. [...] [The system] just doesn't cover everything. And I find it hard to believe that it can."* (P8).

4.3 Means

4.3.1 Effectiveness. Participants see ready availability of relevant information as the greatest value a DST could provide for diversion decisions (explicitly mentioned by 7/8 participants), *"because humans are simply not good at that. But humans are usually good at making decisions based on this filtered information."* (P1). But to be an effective help, the information has to be reliable and as complete as possible. 7/8 participants emphasized that being able to rely upon the system is a fundamental precondition for using and accepting the system; they did not want to ponder over its correctness in the moment of the decision. When we probed for participants' thoughts on the possibility that the system might not be aware of some information, they reacted sceptically:

> *"Do I have to do that every single time? Or can I really rely on the system to give me a good answer?"* (P7)
> *"If I would use the system with this approach, I wouldn't use it at all. [...] Then it occupies me more than making the decision myself."* (P1)

Interestingly, there seems to be a tension between this rejection of deliberate trust calibration at decision time on the one side and not blindly trusting the system on the other side. Pilots are well aware of the risk of overreliance (explicitly mentioned by 5/8 participants), and say they would use the system critically, e.g. by double checking (6/8) or first making an own decision before consulting the system (5/8): *"I would use the system as a mirror and as positive confirmation for the things we have already thought about. [...] Just like a co-pilot in the end, the possibility of: 'ah, I didn't think of that!'"* (P6). We will discuss this tension further in Section 5.1.3.

While information has to be as complete as possible with respect to the current situation, 6/8 participants stated that it also has to be relevant. Some information like passenger handling need to be tailored to the airline, and other information is only relevant in certain situations, e.g. distance to the next hospital.

4.3.2 Efficiency. Given the time pressure often present in diversions, it is important that pilots can work efficiently with the DST-provided information. Ranking the options like in *Global Suggestions* can be a good way to accelerate pilots' decisions. 7/8 participants expressed positive views about the ranking: *"I could see at a glance, will this work, or will it not?"* (P5). 2/8 pilots would go even further and would like the system to filter out unfeasible options.

Local Hints represent another possible way to increase efficiency by directing pilots' attention to potential risks. 5/8 participants explicitly expressed positive opinions about the hints: *"A restrained AI would definitely also have a chance, and not just 'here, this is my recommendation now', or 'do this now', but in the sense of 'here, these are my concerns and here are my references.'"* (P2).

Apart from intelligent data processing, the user interface needs to be efficient. All participants discussed how the UI could be uncluttered, especially for the *Local Hints* prototype. 3/8 participants also remarked that the criteria definition step in the *Global Suggestions* prototype has to be efficient. Lastly, 2/8 participants pointed out that the DST has to be integrated into the entire workflow including other cockpit systems.

4.3.3 Appropriability. As diversions can be highly complex, it is virtually impossible for designers of DSTs to foresee every possible situation. Care needs to be taken so that pilots can still work with the system, even when the situation falls outside of what designers explicitly considered. This is what *appropriation* is about, the adaptation and adoption of technology by users in ways not expected by designers [14]. We present the following points under the umbrella of *appropriability* since they are all about how pilots can incorporate the DST into their task.

5/8 participants highlighted the importance of *transparency*, where *Local Hints* was apparently perceived as more transparent; 5 pilots lauded the transparency in *Local Hints*, only 2 in *Global Suggestions*. Pilots seemed to be more interested in *what* the reasons are for system evaluations, while none of them asked about *how* the system would work: *"It is transparent and you can see what led to the decision of the system to choose and to recommend this airport or not."* (P3). Another transparency need that participants expressed was to know how reliable the DST information is (Section 4.3.1). For example, P5 would like to have indications about the recency of information: *"Is the data brand-new, or is it half a day old? This also makes a difference."* (P5). P2 would like to have a link to the raw data to understand where the information is coming from: *"You could get an input from the AI, 'this is my thought because it says so here and there,' a link from the AI suggestion, where does it take it from, where does it say so, how is the connection?"* (P2)

Apart from transparency, pilots require *directability*, i.e. the possibility to steer the system according to their current needs. In our concepts, directability was primarily given by the pilot-defined criteria in *Global Suggestions* (positively commented on by 4/8 participants) and to a lesser extent by the pilot-specified emergency type in *Local Hints* (4/8 participants). 3/8 participants wanted to have more control over the criteria shown in *Local Hints*, similar to the controls in *Global Suggestions*, underscoring the importance of directability. Pilots also suggested additional means of directability, like options to hide undesired airports (5/8 participants) or to add an airport that is not in direct vicinity (P6). However, controls must not be overwhelming: 3/8 pilots criticized the complexity of the criteria definition. Controls further have to be optional: 3/8 pilots pointed out that the system should be able to provide value without much configuration, but the controls need to be there so that pilots can steer the system according to their needs.

> *"If you can simply [...] click on it, 'bang, show me the airports', then I think it's great, then it's fast, because anything that takes longer, you just don't have time for in an emergency. If it's not time-critical, [...] you can vary the infrastructure yourself, like runways, conditions, etc., in order to be even more precise."* (P4)

With *Local Hints*, we further probed for a third type of appropriability means. The hints offer *continuous support*, also in normal flight, instead of a one-off interaction in reaction to an emergency. This is meant to address pilots' need to maintain SA and to plan ahead (Section 4.1.2), which 5/8 pilots acknowledged explicitly:

> *"You can also at any time, without actually being in an emergency, [...] constantly think about which airports are reasonable [...]. You don't have to wait. I don't like to wait with my planning until the emergency occurs."* (P5)

One pilot fittingly described it as follows:

> *"The [Local Hints] system as a supporting actor in the background, it's quite good. [...] It's like another person plotting along, who keeps saying 'pay attention, consider this maybe', or a person who in the background decodes all the information and plots along and then makes it available."* (P2)

Through continuous support, pilots can also familiarize with the system during normal flight, which is important for pilots to rely on the system in an emergency (mentioned by 4/8 participants). As diversions are rare, a system only designed for use in an emergency would provide little opportunity for familiarization: *"I think I have roughly experienced four or five diversions in twelve years. And if I only started using such a system once such a situation arises, I don't know if pilots would be open to it."* (P2). As one pilot pointed out, continuous support might also help pilots to better incorporate the information into their decision-making:

> *"When situational awareness is always ensured, like with the [Local Hints] system for example, then all the thoughts that I would think about with the [Global Suggestions] system in the case of a diversion, I would already have them in the back of my mind."* (P3)

Lastly, 4/8 participants brought up that pilots would be trained extensively on the usage of such a system, which also helps to properly incorporate it into their decisions.

4.4 Participants' impressions and preferences

All participants had a positive impression of both prototypes, with 5 preferring *Global Suggestions*, and 2 preferring *Local Hints*. Yet, 5/8 pilots explicitly praised the possibility to plan ahead during normal flight, which only *Local Hints* was designed to support. It appears that most pilots did not recognize that *Global Suggestions* was not designed to do that. 4/8 participants explicitly stated that they did not notice a big difference in functionality and saw the main difference in the visual presentation. The overall preference for *Global Suggestions* therefore mostly had to do with the less cluttered presentation of that prototype and its ranking functionality, which makes good options more apparent. 5/8 participants said they would prefer a combination of both concepts.

5 DISCUSSION

5.1 Takeaways

5.1.1 Decisions are a process, not a point. The NDM approach views decision-making as a process that is much broader than merely the point at which an option is selected, as is the usual focus of HB research and controlled studies on DSTs. In line with the NDM view, pilots do not only have reactive, but also proactive goals and actions. Even the reactive goals and actions are not limited to just selecting an option, but also involve for example a continuous check whether the chosen option is still valid. Neglecting the process character of decision-making in the design of DSTs might lead to systems that are not all that useful to decision makers, as found by other real-world studies on DSTs [37, 62]. Acknowledging this process character on the other hand can reveal support opportunities that one would miss when treating decisions as a point. An example for such an opportunity is to continuously support pilots in maintaining SA and planning ahead, even during normal flight when there is no diversion decision to be made.

5.1.2 DSTs can be more (or less?) than oracles. Writing about technology design in general, Alan Dix noted that "*Instead of designing a system to do the task you can instead design a system so that the task can be done.*" [14] Applied to DSTs, this suggests that there might be more helpful uses for AI than being an "oracle", as termed by Cabitza et al. [7]. In line with this thought, P3 remarked that "*I am confident that I can make similarly good decisions [as the system]*" and sees the value of our concepts more in a reduced workload and increased SA. Moreover, decision recommendations from an "oracle" DST might be hard to incorporate into users' decision-making process, especially when the decision requires much more context than the system considers [4, 37], as would likely be the case with diversions. We explored an alternative to an "oracle" DST with the *Local Hints* concept which continuously gives warnings and alerts about individual data points. The aim of this more fine-grained support compared to ready decision recommendations was to allow pilots to more easily incorporate it into their decision-making, even when the support might be imperfect at times. While more pilots overall preferred *Global Suggestions* over *Local Hints* due to the less cluttered presentation and ranking functionality of the former, most pilots viewed the ideas behind *Local Hints* very favorably.

Given that the implicit goal of AI research is frequently to replicate human capabilities [53], a more restrained design that seems to do less can appear less ambitious than an "oracle" DST. Here again, it is worth quoting Dix: "*Designs that are closed are often more apparently sophisticated, because they may do more for the user, but ultimately do not allow the users to do more for themselves.*" [14] We see our results as motivation for further investigations into more restrained DST designs like our *Local Hints* concept. Examples from prior work include the concept of *Unremarkable AI* by Yang et al. [61] or the *interaction-as-commentary* paradigm described by van Berkel et al. [57].

5.1.3 Appropriation leads to resilience. Diversions are highly complex, but intelligent systems currently often cannot deal with every facet of this real-world complexity, and some would argue that they never will [59]. The challenge is therefore to design the joint human-machine system to be resilient, even though the intelligent system on its own is brittle. With DSTs, the usual focus to this end is on trust calibration, which was not well received by our participants. We therefore propose to frame the challenge of resilient human-machine systems around appropriation.

The lens of appropriation might help to resolve the tension that we observed in our interviews between pilots' rejection of deliberate trust calibration on the one hand and not blindly trusting the system on the other hand: Pilots are aware that a DST cannot cover every eventuality, but they do not want their task to be primarily about questioning the correctness of the system. Rather, they want the system to complement their decision-making, i.e. they want to be able to appropriate the DST. There have been similar findings for instance for clinical decision-making, where clinicians would not want to calibrate their trust in a DST for every decision [32].

In this view, transparency may not primarily be about trust calibration. This is not to say that trust calibration should be dismissed; whether users should adopt a system output or not is definitely important, but it might be a too simplistic perspective for real-world usage. Rather, transparency can be seen as a means for appropriation. This includes considering how users might appropriate explanations, as Ehsan et al. have found [17]. But it also goes further, with the question being how transparency can help users to appropriate the DST—or more generally, the intelligent system—as a whole. For instance, applied to intelligent systems, two of Dix's well-known guidelines for appropriation [14], *provide visibility* and *expose intentions*, immediately call transparency to mind. As a negative example in this regard, Blomberg et al. [4] provide a case study from enterprise analytics, where a sales team was unable to appropriate a machine learning model due to the lack of interpretability. The project was abandoned despite the good performance of the model.

Our results further indicate that transparency is not the only lever to design for appropriation in intelligent systems. Directability is another means that is often given much less attention than transparency, but might be just as important. If directability is discussed, it is often in the context of *interactive machine learning* [1, 15]. However, directability is not only about directing what a model learns, but also about steering the system according to the current user intentions. Sometimes, the intention might be highly situation-specific and not a general pattern that the model should learn. In our concepts, directability was primarily given by the possibility for pilots to specify diversion criteria in *Global Suggestions*. An example from literature is the content-based medical image retrieval system by Cai et al. [8], which included tools for pathologists to direct the system output according to their intentions. Interestingly, the authors found that pathologists appropriated these tools to understand how the algorithm worked and whether surprising outputs resulted from algorithm errors or their own oversights [8].

An even less researched appropriation means is continuous support, as exemplified in our *Local Hints* concept. Participants' remarks hinted at how continuous support might facilitate appropriation, e.g. through increased familiarity with the system, or by already being mentally engaged with the system output should an emergency occur. Further research should investigate whether these conjectures can be confirmed empirically. Chiou and Lee at least argue that the particular sequence of human-system interactions has a big influence on the usage of intelligent systems [10]. Taken together, while trust calibration is important, the means to allow humans to work effectively and resiliently with intelligent systems are much broader through the lens of appropriation.

5.2 Limitations

Our results need to be interpreted in light of following limitations: For one, given the difficulty of recruiting professional pilots for extensive interviews, our sample size was still relatively small with only eight pilots. Moreover, we only had participants from German airlines. This might have an influence, as there are regional differences in pilots' working culture, e.g. in how stringently pilots adhere to procedures or how willingly first officers speak up against their captain [28]. Furthermore, as the interviews were based on demonstrations of click dummy prototypes, pilots were not able to test our concepts. This led to contradictory opinions, e.g. about which concept would be more efficient under time pressure. Our work serves as an exploration of possible design opportunities and requirements which have to be evaluated empirically in future work.

6 SUMMARY AND OUTLOOK

We designed and confronted pilots with two low-fidelity prototypes to explore how an intelligent system might support highly complex diversion decisions. We found that there are more opportunities to support pilots than just suggesting diversion options, e.g. supporting their need to always maintain situation awareness, also during normal flight. As for design requirements, we suggest that a useful lens is how pilots can appropriate the intelligent system for their decision-making. Appropriation might not only be enabled by transparency, but also through directability and continuous support. We see our results as a call for future research on intelligent decision support tools to expand the view beyond trust calibration at the actual point of decision. This likely requires more investigations of real-world decision environments to understand how decision-making goes beyond an isolated selection between options.

ACKNOWLEDGMENTS

In memory of our dear colleague and mentor, Heinrich Hußmann, who sadly passed away much too soon.

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

We thank our participants for their time and valuable input. We further thank Dr. Christoph Vernaleken, Felix Ehrentraut, and Alexander Hennrich, who were of great help for us to get to know the field of aviation in general and diversions in particular.

REFERENCES

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

[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] Charles E. Billings. 1996. *Human-centered aviation automation: principles and guidelines.* Technical Report NASA-TM-110381, A-961056, NAS 1.15:110381. NASA Ames Research Center, Moffett Field, CA, USA. 222 pages. https://ntrs.nasa.gov/citations/19960016374 .

[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] 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 .

[6] 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 .

[7] 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 .

[8] 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 .

[9] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, Scotland, UK, 559:1–559:12. https://doi.org/10.1145/3290605.3300789 .

[10] Erin K. Chiou and John D. Lee. 2021. Trusting automation: designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society* Online ahead of print (April 2021), 1–29. https://doi.org/10.1177/00187208211009995 .

[11] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, College Station, TX, USA, 307–317. https://doi.org/10.1145/3397481.3450644 .

[12] Mary L. Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*. Chicago, IL, USA, 33:1–33:6. https://doi.org/10.2514/6.2004-6313 .

[13] 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 .

[14] 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

[15] 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 .

[16] EASA. 2020. *Artificial intelligence roadmap: a human-centric approach to AI in aviation.* Technical Report. European Union Aviation Safety Agency (EASA). 33 pages. https://www.easa.europa.eu/sites/default/files/dfu/EASA-AI-Roadmap-v1.0.pdf .

[17] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I.-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The who in explainable AI: how AI background shapes perceptions of AI explanations. https://doi.org/10.48550/arXiv.2107.13509 .

[18] Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors* 37, 1 (March 1995), 32–64. https://doi.org/10.1518/001872095779049543 .

[19] Mica R. Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59, 1 (Feb. 2017), 5–27. https://doi.org/10.1177/

0018720816681350 .

[20] Claudia Fernández Henning. 2021. *FOR-DEC and beyond – conceptual HMI design for a diversion assistance system.* Bachelor Thesis. Technische Hochschule Ingolstadt, Ingolstadt, Germany.

[21] Therese Fessenden. 2019. *The diverge-and-converge technique for UX workshops.* Nielsen Norman Group. Retrieved 2023-01-23 from https://www.nngroup.com/articles/diverge-converge/

[22] Ken Funk, Beth Lyall, Jennifer Wilson, Rebekah Vint, Mary Niemczyk, Candy Suroteguh, and Griffith Owen. 1999. Flight deck automation issues. *The International Journal of Aviation Psychology* 9, 2 (April 1999), 109–123. https://doi.org/10.1207/s15327108ijap0902_2 .

[23] Thomas Gilovich and Dale Griffin. 2002. Introduction – heuristics and biases: then and now. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, Dale Griffin, Daniel Kahneman, and Thomas Gilovich (Eds.). Cambridge University Press, Cambridge, 1–18. https://doi.org/10.1017/CBO9780511808098.002

[24] Cleotilde Gonzalez and Joachim Meyer. 2016. Integrating trends in decision-making research. *Journal of Cognitive Engineering and Decision Making* 10, 2 (June 2016), 120–122. https://doi.org/10.1177/1555343416655256 .

[25] 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 .

[26] Francisco Gutiérrez, Nyi Nyi Htun, Vero Vanden Abeele, Robin De Croon, and Katrien Verbert. 2022. Explaining call recommendations in nursing homes: a user-centered design approach for interacting with knowledge-based health decision support systems. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM, Helsinki, Finland, 162–172. https://doi.org/10.1145/3490099.3511158

[27] 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 .

[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] EAAI HLG. 2020. *Fly AI report - demystifying and accelerating AI in Aviation/ATM.* Technical Report. European Aviation High Level Group on AI (EAAI HLG). 76 pages. https://www.eurocontrol.int/publication/fly-ai-report .

[30] Karen Holtzblatt and Hugh Beyer. 2017. The affinity diagram. In *Contextual Design: Design for Life* (2 ed.). Morgan Kaufmann, Boston, 127–146. https://doi.org/10.1016/B978-0-12-800894-2.00006-5

[31] 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)*. Dublin, Ireland, 17–23. https://elib.dlr.de/27044/ .

[32] 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 .

[33] 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 .

[34] Daniel Kahneman. 2011. *Thinking, fast and slow* (1st ed.). Farrar, Straus and Giroux, New York. .

[35] Daniel Kahneman and Gary Klein. 2009. Conditions for intuitive expertise: A failure to disagree. *American Psychologist* 64, 6 (2009), 515–526. https://doi.org/10.1037/a0016755 .

[36] 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 .

[37] 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 .

[38] Gary A. Klein. 1993. A recognition-primed decision (RPD) model of rapid decision making. In *Decision making in action: Models and methods*, Gary A. Klein, Judith Orasanu, Roberta Calderwood, and Caroline E. Zsambok (Eds.). Ablex Publishing, Westport, CT, US, 138–147.

[39] Gary A. Klein. 2008. Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 456–460. https://doi.org/10.1518/001872008X288385 .

[40] Gary A. Klein, Roberta Calderwood, and Donald MacGregor. 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 3 (May 1989), 462–472. https://doi.org/10.1109/21.31053 .

[41] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-AI decision making: a survey of empirical studies. https://doi.org/10.48550/arXiv.2112.11471 .

[42] 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

[43] Raanan Lipshitz, Gary Klein, Judith Orasanu, and Eduardo Salas. 2001. Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making* 14, 5 (Dec. 2001), 331–352. https://doi.org/10.1002/bdm.381 .

[44] 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 .

[45] 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 .

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

[47] Austrian Federal Office of Transport. 2006. *Untersuchungsbericht: Flugunfall mit dem Motorflugzeug Type Airbus A310 am 12. Juli 2000 am Flughafen Wien-Schwechat, Niederösterreich.* Investigation Report GZ. 85.007/0001-FUS/2006. Austrian Federal Office of Transport, Vienna, Austria. 89 pages. https://reports.aviation-safety.net/2000/20000712-0_A310_D-AHLB.pdf

[48] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, LA, USA, 568:1–568:9. https://doi.org/10.1145/3491102.3502104 .

[49] Emilie Roth, Devorah Klein, and Katie Ernst. 2021. *Aviation decision making and situation awareness study: decision making literature review.* Contract Report USAARL-TECH-CR–2022-17. Applied Decision Science, Cincinnati, OH, USA. 58 pages. .

[50] Nadine B. Sarter, David D. Woods, and Charles E. Billings. 1997. Automation surprises. In *Handbook of Human Factors & Ergonomics* (2 ed.), Gavriel Salvendy (Ed.). John Wiley & Sons, 1926–1943. .

[51] 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 .

[52] 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 .

[53] 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 .

[54] 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 .

[55] Henning Soll, Solveig Proske, Gesine Hofinger, and Gunnar Steinhardt. 2016. Decision-making tools for aeronautical teams: FOR-DEC and beyond. *Aviation Psychology and Applied Human Factors* 6, 2 (Sept. 2016), 101–112. https://doi.org/10.1027/2192-0923/a000099

[56] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 4157 (Sept. 1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 .

[57] Niels van Berkel, Mikael B. Skov, and Jesper Kjeldskov. 2021. Human-AI interaction: intermittent, continuous, and proactive. *Interactions* 28, 6 (Nov. 2021), 67–71. https://doi.org/10.1145/3486941 .

[58] 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 .

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

1177/1555343416653562 .

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

[61] 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 .

[62] 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 .

[63] 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 .

[64] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hußmann. 2021. Pilot attitudes toward AI in the cockpit: implications for design. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. IEEE, Magdeburg, Germany, 159–164. https://doi.org/10.1109/ICHMS53169.2021.9582448 .