BRAID
Bridging Responsible AI Divides

July 2024

Explainability in Expert Contexts

Challenges and Limitations in Supporting Domain Experts in AI-driven Decision-making

Auste Simkute

This fellowship ran from January-May 2023 as part of BRAID.

BRAID is a UK-wide programme dedicated to integrating Arts and Humanities research more fully into the Responsible AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI. Funded by the Arts and Humanities Research Council (AHRC), BRAID represents AHRC's major investment in enabling responsible AI in the UK. The Programme runs from 2022 to 2028. Working in partnership with the Ada Lovelace Institute and BBC, BRAID supports a network of interdisciplinary researchers and partnering organisations through the delivery of funding calls, community building events, and a series of programmed activities. Funding reference: Arts and Humanities Research Council grant number AH/X007146/1.

Learn more at **www.braiduk.org**

This research was supported via UK Research and Innovation by the R&D Science and Analysis Programme at the Department for Culture, Media & Sport. Any primary research, subsequent findings or recommendations do not represent Government views or policy and are produced according to research ethics, quality assurance, and academic independence.

To request an alternative format of this report please email **braid@ed.ac.uk**.

Table of Contents

Abstract

Advanced artificial intelligence (AI) techniques are increasingly used to support often sensitive decisions in the public sector. AI in expert contexts is intended to enhance expertise and conserve expert time by taking over mundane tasks. The goal is an effective expert–AI collaboration. However, these decision-support technologies can overwhelm domain experts, such as social workers, healthcare workers, and recruiters. The introduction of AI often disrupts experts' ability to make decisions in their preferred way and interrupts their workflow. This can result in low adoption of AI systems, as experts report feeling burdened by them and impeded in applying their expert skills. In addition, experts often cannot understand how these systems work and, as a result, either over-rely on or distrust AI-provided recommendations.

This can lead to experts either becoming overly reliant on AI systems or refusing to use them completely. Without meaningful expert input, sensitive decisions, such as who will receive refugee status or qualify for unemployment benefits, are left, almost in their entirety, to automation. A promising solution is to explain the AI processes and recommendations. Explainability is a set of techniques that provide human-understandable information about a system's behaviour, processes, and outputs. However, current explainability methods are often ineffective, and even misleading, when applied in expert contexts.

Based on the literature review and a preliminary contextual enquiry study with science experts and AI developers, this report provides an overview of the challenges faced when introducing decision-support systems in practice. It explores critical blockages for effective human–AI collaborations and discusses potential solutions. It also discusses the role of explainability in supporting experts and outlines recommendations for how explanations could be made more effective and usable.

Key Takeaways

The public sector increasingly relies on artificial intelligence (AI) driven systems to cope with vast amounts of available data and improve decision-making in various contexts, such as healthcare and social work. However, experts often need help to make informed judgements based on AI-generated outputs and lose their agency and ability to apply their expertise. Explainable AI techniques could provide a solution, but they are often ineffective, and even misleading, when used in expert contexts.

Based on the literature review and a preliminary contextual enquiry study with science experts and AI developers, this report provides an overview of challenges faced when introducing AI-driven decision support systems (DSSs) in practice. It discusses the role of explainability in supporting experts and outlines recommendations for how explanations could enable effective human–AI collaboration.

Main Findings

- Introducing DSSs disrupts experts' decision-making strategies and interrupts their workflows, limiting their ability to apply expertise.
- Loss of agency and limited opportunities to make AI-independent decisions can result in expert de-skilling.
- Explainability can create an illusion of trustworthiness, resulting in experts becoming overly reliant on DSSs.
- Explanations that are not tailored to a specific domain and expert needs will likely be ignored.
- Explanations the evaluation of which requires technical knowledge are seen as burdensome and frustrating by domain experts.

Recommendations for Effective AI-expert Collaboration

- There should be an adjustment period when DSS is first introduced.
- Domain experts should be involved in technology planning, development, design, and implementation.
- Domain experts should be able to continuously collaborate with the developers and provide feedback about the system's performance, usability, and contextual fit.

Recommendations for Effective Explainability

- Explainability should be accessible to domain experts independently of their computational and data-science understanding.
- Explainability should be meaningful to domain experts (e.g. using domain-specific terminology) and fit within their workflows.
- Explainability should be interactive, support flexible information search, and provide domain-specific context through examples.
- Explainability should use techniques, such as cognitive forcing (e.g. asking a user to explain their reasoning), to encourage experts to engage with the explanations mindfully and to remain motivated.

Executive Summary

Context

The public sector increasingly relies on artificial intelligence (AI) to inform decision-making across various domains, including policing, healthcare, social work, and immigration services. AI decision support systems (DSSs) can process large amounts of data (1) and generate outputs, such as predictions of medical diagnoses (2) or potential outcomes of a visa application (3). AI support could make processes within the public sector not only more efficient but also fairer by reducing the potential for human biases (4, 5).

However, AI-driven systems lack contextual sensitivity and cannot account for unique cases. They can also be trained on biased or incomplete data. Given that most of the decisions are highly sensitive, it is crucial that domain experts (e.g. social workers) maintain agency when making AI-supported decisions. Ideally, AI would automate mundane, repetitive tasks and allow experts to focus on higher-level and creative ones (6). Unfortunately, domain experts often cannot understand and evaluate whether they should trust AI systems and their generated outputs (7).

This report provides a broad overview of challenges faced when DSSs inform decision-making. It explores critical blockages for effective expert–AI collaborations and discusses potential solutions. It also considers the role of explainability in supporting experts and outlines recommendations for how explanations could be made more effective and usable in each expert context.

Methodology

This report is based on a systematic literature review of research publications that explore explainability in AI-supported decision-making. To better understand the fundamental information needs of experts, a further literature review was conducted exploring human factor engineering research literature on expert decision-making with the support of automation or AI systems.

The recommendations were also informed by a contextual enquiry study with AI developers and expert scientists using AI systems to inform their research in areas of biology and medicinal chemistry. These interviews were focused on uncovering the blockages for effective AI systems implementation in expert contexts.

Key Findings

Reviewed studies showed that AI support does not improve experts' decision-making accuracy. Expert–AI collaboration is often less effective than either AI or an expert working individually (8). AI-driven DSSs can disrupt experts' ability to use their knowledge and skills. Unfamiliar and incomprehensible systems might also cause frustration among experts, as attending to them adds to their already busy workload (9).

Expert–AI Collaboration Challenges:

- Introducing DSSs disrupts experts' decision-making strategies and interrupts their workflows, limiting their ability to apply expertise.
- AI systems can be incomprehensible due to their complex and opaque algorithmic nature. Domain experts might need a more technical background or training to interpret AI-generated outputs correctly and assess their trustworthiness.

Loss of agency and limited opportunities to make AI-independent decisionscan result in expert de-skilling.

To solve these issues, AI-generated outputs are often supported with explanations (10, 11, 12) – technical solutions also called 'eXplainable AI' (XAI). These explanations should give users insights into how an AI output was generated, such as the most significant factors (11, 13). However, most explainability approaches are ineffective or increase the risk of over-reliance when applied in expert contexts.

Limitations of Explainability in Expert Contexts:

- If explanations do not fit the context, domain experts will likely reject them. For example, if an overly informative explanation is presented in a high-time-pressure context, experts will not have enough time to attend to it.
- If explanations are not domain specific, they will likely be overlooked or seen as not applicable by experts. For example, experts will find applying it in each scenario challenging if an explanation does not include domain-specific terminology.
- If explanations are too simplistic and repeat experts' domain knowledge, they are considered redundant by experts.
- If explanations are overly technical and require data science or other skills that are not domain-specific, experts see them as burdening and frustrating.
- Explainability can create an illusion of trustworthiness, resulting in experts over-relying on DSSs.

Recommendations for Effective AI–Expert Collaboration

Findings from the literature review and contextual enquiry study were used to shape recommendations that could help enable effective collaboration between domain experts and AI systems. First, several steps must be taken to lay the foundations for explainability. Meeting these base requirements before attempting to implement explainability approaches could improve the chances that explainability will be effective.

- **There should be an adjustment period when a new AI system is introduced.** Experts should experience the system through explorative exercises. While exploring, they should have easy access to support from the AI team, designated support staff, or experienced peers. Active collaboration between experts and developers should be accommodated during this stage.

- **Experts should be involved in technology planning, development, design, and implementation**. Continuous collaboration between experts and developers could ensure that a new AI system accommodates expert needs, adjusts their expectations towards the technology, and does not interfere with their workflows.
- **There should be straightforward ways for experts to give feedback about the system's performance or suitability**. Experts should be encouraged to submit their comments and suggestions, and be informed of changes influenced by their feedback.

Explainability Recommendations:

- **Use interactive interface design elements.** The initial exploration of the system could be enhanced using interactive elements, such as the ability to simulate multiple potential outputs by interacting with contributing features.
- **Ensure that explainability is available and visible.** Explainability should be accessible to domain experts. They should know how to access and interpret these explanations (14).
- **Avoid overly technical explanations.** Explainability should be accessible to domain experts irrespective of their computational and data science understanding (15).
- **Tailor explanations to a domain and task.** Explainability should be meaningful to experts (e.g., using domain-specific terminology) and fit within their workflows (16).
- **Align explainability with expert decision-making strategies.** Explainability should support flexible information search and outcome comparisons and provide domain-specific context through examples.
- **Use engaging design elements.** Use techniques, such as cognitive forcing (e.g., asking them to explain their decision), to encourage experts to attend to the explanations mindfully and remain motivated to learn about the system (17).

Further Questions

There is a need to understand the long-term effects of DSS use by domain experts – how these technologies will affect expertise developments and de-skilling can only be extrapolated. However, it remains unknown how experts will adapt over time, whether they will transfer saved cognitive resources to strengthen non-automated skills, or develop new expert skills. Understanding how techniques such as explainability could help them do this is essential. This report is an initial step in this direction. However, more effort must be put into ensuring that experts do not become passive observers of AI work. This is particularly important considering the rapid development of large language models and generative AI more broadly.

This review revealed that for AI–expert collaboration and explainability to be effective, there is a need for developers and experts to collaborate. A question for future research is how to overcome the initial communication blockages and frustrations to achieve a collaborative stage where both teams align their communication styles. Without directly involving experts in various DSS development and implementation stages, building usable technologies and motivating experts to learn about and adopt them will be challenging. This could lead to wasting resources on technology developments and prevent experts from benefiting from DSS.

1. Introduction

Box 1. Definitions

Artificial intelligence (AI) – in this report, it is an umbrella term for various algorithmic and statistical tools, including machine learning models (11).

Decision support systems (DSSs) – AI-driven systems that generate data-based outputs that decision-makers use to make informed judgements (18).

Explainability – a set of algorithmic techniques intended to reveal a system's strengths and weaknesses and make its reasoning comprehensible to humans (19, 11).

Explanation – an interface between a human and an AI system that accurately describes the algorithm's logic and outcome and is comprehensible to a human (19).

Expert – a trained professional with experience in a particular domain whose expertise results from extensive and deliberate practice (20).

Governments increasingly rely on artificial intelligence (AI) tools and systems to manage large amounts of data (1). As a result, AI decision support systems (DSSs) are increasingly prevalent in the public sector (21, 22). DSSs generate outputs (e.g. predictions, recommendations) that inform expert judgement in a range of domains, including policing, healthcare sector, recruitment, and social and immigration services. (For more information, see Review into bias in algorithmic decision-making (23)).

This report reviews recent evidence on the challenges and risks associated with experts making decisions based on AI-generated recommendations, produced in a way that is not understandable to them, and does not allow them to effectively evaluate their trustworthiness. It identifies critical barriers, such as explanations not fitting within experts' workflows, that are preventing the practical application of available explainability methods. It also recommends approaches that could make AI recommendations and their explanations more accessible and understandable to domain experts.

Box 2. DSS examples

Example 1: Employment. DSSs can profile an unemployed individual's case by calculating their risk score based on variables such as education, age, gender, type of housing, etc. Based on these scores, the DSS determines the kind of programme for which the applicant is eligible (e.g., job placement, vocational training, apprenticeship, activation allowance) and warns if someone is at risk of long-term unemployment (24, 25).

Example 2: Immigration. DSSs can screen individual applications for a visa, refugee status etc., and make recommendations to the immigration worker concerning whether a person should be allowed to enter a country and which cases should be evaluated with additional scrutiny (3, 26).

Example 3: Diagnostic medicine. DSSs can help cellular pathologists make diagnostic decisions based on biopsy results. AI systems can screen out typical results and inform physicians of the atypical ones. This can speed up the diagnostic process and reduce pathologists' workload (2).

1.1 Benefits of Decision Support Systems

DSSs can speed up decision-making and help in situations where many actions need to be taken on a large scale within a limited time (18). AI-driven systems can also reduce the potential for human decision-making biases by providing a more systematic approach (27, 28). In addition, they can outperform human capabilities, especially in rule-based and repetitive tasks (29). In the public sector, DSSs can automate mundane tasks, such as calculating tax returns, or more nuanced tasks, such as granting refugee status in a country (30). The premise for AI-driven decision-making is that better information should lead to better decision-making (31).

Box 3. The Benefits of DSSs

- **Efficiency.** DSSs can undertake high-volume, repetitive tasks and data-heavy workflows, freeing up experts' valuable time and processing a vast amount of data (29, 32).
- **Effectiveness.** DSSs can enhance the quality of services for citizens, e.g., they can speed up decision-making and turnaround time of an outcome (29, 33, 34, 5).
- **Fairness.** DSSs can increase decision-making consistency and reduce human biases in decisions (35, 36, 37).
- **Consistency.** DSSs can be highly organised, thorough, and systematic (5).

1.2 Pitfalls of Decision Support Systems

However, using DSSs can lead to inaccurate decisions that can be particularly costly in high-risk and sensitive domains – for more information about bias in algorithmic decision-making, refer to the Review into bias in algorithmic decision-making (23) and Barocas and Selbst (38). Furthermore, using AI-driven systems in the public sector also poses a risk to values such as accountability, transparency, equality, privacy and security, sustainability, and interoperability – for more information on the broader impact of AI-supported decision-making, see Brauneis and Goodman (39), Nair et al. (28), Kankanhalli et al. (34), and Ehsan et al. (41). Without meaningful human input, algorithmic unfairness might remain unrecognised until a targeted investigation is conducted (42) and might lead to the replication and even amplification of existing biases in society (43, 44).

Box 4. The Weaknesses of DSSs

- **Opaqueness.** Complex AI systems are ‘black boxes’, meaning their inner processes are opaque and incomprehensible to humans (19). The lack of transparency means these systems can only be inspected if explainability techniques are applied (45).
- **Unfairness.** AI-driven tools are susceptible to errors due to biased or incomplete datasets (30, 46). This can lead to societal biases being reflected in DSS outputs (38).
- **Accountability.** It can sometimes be made unclear who should be held accountable in case of an error (47, 48). For example, experts might be held responsible for the outcomes even if they have little agency to inspect and override DSS outputs (49).
- **Oversimplification.** Fair and valid assessments often require detailed data, preserving contextual information. However, DSSs are mostly not sensitive to context and might ignore essential factors that should be considered (44).
- **Implied causality.** AI-driven systems are trained to find statistical correlations. Such correlations might or might not be caused by causal relationships. For example, a DSS might suggest a meaningful connection even if no causal relationship exists (44).

2. Research Method

This report is based on a systematic literature review of research publications exploring explainability in AI-supported decision-making. Most of the reviewed explainability studies involving domain experts as participants were conducted within a healthcare domain. Other expert-focused studies included social work, recruitment, and immigration. Relevant search terms were determined by examining the variations in terminology used in the most influential articles in the field of explainable AI and closely related to the main research questions. The list of papers was manually filtered using set inclusion criteria and excluding publications not investigating the expert/decision-maker as a stakeholder.

To better understand the fundamental information needs of experts, a further literature review was conducted exploring the human factor engineering research literature on expert decision-making with the support of automation in contexts such as air traffic control, aviation, and intelligence. The list of papers was manually filtered using set inclusion criteria, excluding publications not investigating human psychology aspects in the decision-making context or findings that were not relevant and transferable in the algorithmic decision-making context.

In addition, the findings of a contextual enquiry study with AI developers and scientists using their developed software were also used to inform the report. This study involved five in-depth interviews with members of the AI team and five observations with in-depth interviews with experts using AI-driven technologies to analyse their data.

3. DSSs in Expert Contexts

Sensitive decisions, such as who will receive refugee status in a country, are highly complex and discretionary (3, 50). These decisions require a human to show empathy, consider unusual circumstances, and notice salient factors not reflected in training data (44, 2). The fundamental goal of a DSS is to augment the decision-making process rather than fully automating it (51). Ideally, AI would automate mundane, repetitive tasks and allow experts to focus on higher-level and creative ones (6). However, humans should stay in control of making the final decision, while automation is used to aid it. Having a human oversee the workings of AI has been shown to be an effective way to reduce errors in medicine (52) and the legal sector (53). Experts also express a need to maintain a sense of control and autonomy in decision-making (54, 55).

Box 5. Expert Skills That Could Be Enhanced by DSSs

- Building narratives to integrate available information and simulate different potential outcomes (44).
- Deducting and verifying the potential options/outcomes (44, 18).
- Recognising out-of-the-ordinary patterns and unique cases (44, 18).
- Integrating knowledge from different sources (2).
- Making initial insights and impressions of the situation and intuitively linking that to potential outcomes/reducing available options (2).
- Drawing up conclusions despite incomplete information and uncertainty (51, 6, 56).
- Being more creative in medical problem-solving (52).
- Making more rational decisions (53).
- Being more mindful of one's own biases (53).

3.1 DSSs in Expert Contexts: Challenges and Limitations

Simply providing recommendations does not improve the accuracy of experts' decisions (56, 18, 57, 58). The use of DSSs can even result in poorer performance compared with a human or an AI system working alone (59, 60). Furthermore, DSS recommendations can make humans doubt their expertise, even when they are correct (61), and over-rely on algorithmic solutions (62). The challenges preventing domain experts' effective use of DSSs must be understood in order to design usable and practical decision support systems.

Box 6. DSS Challenges and Limitations

- Introducing DSSs disrupts experts' decision-making strategies and interrupts their workflows, limiting their ability to apply expertise (56, 57).
- AI systems can be incomprehensible due to their complex and opaque algorithmic nature. Domain experts might also need a more technical background or training to interpret AI-generated outputs correctly and judge whether to trust them (70, 112).
- Less experienced decision-makers are more likely to over-rely on DSSs (68, 69, 70).
- Loss of agency and limited opportunities to make AI-independent decisions can result in expert de-skilling (75, 76).

3.1.1 Changes in the Decision-making Process Prevents Experts From Using Their Expertise

The introduction of DSSs disrupts experts' workflows and changes how they make decisions (63, 64). Without the support of AI, experts can intuitively spot irregularities in data or notice patterns that initially seem insignificant and are unlikely to be picked up by a DSS (65). However, when new factors, such as algorithmic support, are introduced, they cannot apply these skills in the same way (66). Disrupted decision-making leaves experts feeling restrained by the static nature of the DSS predictions (67) and unable to exercise skills they gained while working without algorithmic support (68). Moreover, being spoon-fed recommendations without additional information can be frustrating and demotivating (69, 64, 67). The feeling of confusion forces experts to surrender to their old decision-making methods (even if less effective), for example, by manually searching for information (36).

3.1.2. Disruption of Experts' Workflows Prevents Them From Benefiting from DSSs

Failure to appreciate the context in which decisions are typically made without algorithmic support is one of the reasons why predictive systems fail in practice (49). Poor contextual fit means decision-makers might feel limited and resist relying on a system's predictions (70, 67). They might also lack the means or time to make an informed decision (49). Interviews with public sector workers showed that the way users interact with algorithms, and whether they rely on them, might depend on how well the system fits with their natural workflow and organisational context (71). The disruption to the decision-making workflow prevents experts from using decision-making strategies learned with experience (66). Subsequently, experts, when introduced to the DSS, are likely to rely on their common sense or heuristics, usually searching for aspects confirming their intuition and failing to notice errors (72). As a result, less experienced decision-makers are more likely to rely on AI-driven systems. However, introducing new technologies can inhibit their skills, making them less adaptive and more passive, pigeonholing or disconnecting them from how they would normally prefer to analyse the data (73).

3.1.3 Varying Levels of Expertise Influence Trust in the System

Decision-makers in expert contexts often have different levels of domain knowledge. This level of expertise can influence their acceptance of AI-enabled systems, their outputs, and their initial trust when starting to use them (74). Novice users often struggle to calibrate their trust based on the observed DSS performance and over-rely on algorithmic advice (75, 76, 77, 16). Unjustified novice acceptance of technologies has been observed in the radiology sector (78) and among immigration workers (50). Experienced decision-makers are often more sceptical about new technologies in their expertise context (74). Their perception of system accuracy is also susceptible to first impressions. Observing errors early in the process can lead to experts rejecting algorithmic systems, whereas experiencing high system reliability can lead to future bias towards automation (74). Overreliance on DSSs could be harmful if the AI-generated advice is inaccurate (79, 80).

3.1.4 Public Sector Workers Have Fewer Opportunities to Develop Their Domain Expertise

When a DSS is introduced in a decision-making context, public sector workers do not have the same exposure to naturalistic decision-making. It means they rely on AI recommendations but don't have to analyse, gather, or process information themselves (15). This change in decision-making can negatively affect expertise development and even lead to the loss of expertise (de-skilling) (81). From a long-term perspective, this can mean that valuable human input and unique expert skills, such as intuition and pattern recognition – that AI cannot replace – would be potentially lost (82). Furthermore, this could lead to the decision-makers apathy towards AI-supported systems (83). When an expert is resigned to the fact that they cannot add value to the computer guidance, they lose motivation and are less likely to learn from the past and build expertise (73). This emphasises the importance of proper expert education in building quality experiences so that humans remain 'in the loop'. Such instruction includes formalised training on new concepts and training simulations on past events.

3.2 DSSs in Expert Contexts: Recommendations

Box 7. DSS Recommendations

- There should be an adjustment period when a DSS is first introduced.
- Experts should be involved in technology planning, development, design, and implementation.
- Experts should be able to collaborate with the developers and provide feedback about the system's performance, usability, and contextual fit.

3.2.1 Introduce a Transition Period

Domain experts should have time to adjust to the new algorithmic system when it is first introduced. During this stage, experts should have access to the information about the system. They should also be able to ask questions about it directly (84). Ideally, they could experiment and learn about the DSS through practice (85). During this period, experts should be able to provide feedback about the system and how it fits within their workflow to its designers and developers and work collaboratively towards improving it (86). Having time assigned to observe a system's performance can influence how users will interact with it in the future (74). Without a transition period, experienced workers are more likely to reject these systems, even for non-objective reasons (75). On the other hand, novices that do not have the assigned time to explore the DSS might over-rely on their outputs (88). Guiding users' understanding of AI capabilities and limitations has been shown to be effective in building appropriate trust (89).

Moreover, experts interviewed in the contextual enquiry study also reported that the initial phase of using the system is crucial. They reported being quick to dismiss it at this phase, and being unwilling to invest their time in learning about it further, if the first impression was unsatisfactory. However, if they received support and did not have to find out themselves how to use it, they felt relieved and were willing to put effort into exploring the system.

3.2.2 Place an Expert in the Centre of the Design and Development of AI-supported Systems

Experts should be involved in technology planning, development, design, and implementation stages, to ensure that the new AI system accommodates their needs and is compatible with their workflow (90, 91). Involving experts in various stages of the design is effective with rehabilitation therapists (36), prostate cancer pathologists (89), and cardiologists (92). Consistent communication among professionals, developers, and multidisciplinary researchers could also help to manage experts' expectations for the DSS and motivate them to invest time and effort into learning about the system (73). This collaborative effort has been shown to effectively promote the latest scientific ideas and technologies in experts' operations (73).

Moreover, the interviews with experts and AI developers during the contextual enquiry study revealed that both sides seek better communication. AI team members wanted to understand the task-specific aspects better, and they wanted to learn directly from domain experts. They also wanted to set realistic expectations of what could be achieved using AI tools. On the other hand, domain experts wanted to learn more about the tools they use from the team that developed them. They also wanted to communicate their needs and ask questions when they experienced issues using the software. However, this communication had to be effortfully put in place and even forced initially, as both sides avoided actively initiating contact. Even when they did, communication was too infrequent to be effective. One of the success stories from the interviews confirmed that. Weekly meetings helped experts and AI team members align their ways of communicating and gain confidence to express their needs and voice concerns openly. However, both sides were frustrated during the first few meetings and reported being unable to understand each other. This communication had to be continuous for effective collaboration to be established.

3.2.3 Promote Expertise Development Through Feedback

Development of expertise can be facilitated by receiving feedback that informs experts about their performance (94). The design of the DSS interface should enable practice and allow users to obtain feedback. Feedback shapes a better understanding of contributing features and improves expert judgements (95). Learning from feedback is effective when users are allowed to interact with an automated system and its provided information rather than passively observing it (96, 97, 98). Explaining ADSS outputs with feedback can improve the self-awareness of decision-makers (99) and prevent overconfidence (64). Experts interviewed in the contextual enquiry study also reported the lack of feedback they received as frustrating. They were unwilling to use a system if they did not know if their actions were correct.

3.2.4 Promote Expertise Development Through Interactive Interface Design Features

Simulating AI and automation-generated outputs have been shown to aid expertise development and build a better understanding of the domain (100, 101, 102). Interactive features allow decision-makers to analyse and manipulate different outcomes (103). The ability to make changes to the model by, for example, adding class labels and tuning the classifier's parameters also fosters learning (104). Interactive visualisations are particularly effective (105). Being able to manipulate visual elements directly allows for interactions that are easier to interpret than other types of data display (105). Interactively experimenting with visual features can also help to identify which data items are affected by and related to specific features (105). The ability to make even minimal alterations has been shown to give users a sense of control and increase trust in the system (106).

3.2.5 Promote Expertise Development Through Engaging Interface Design Features

Expertise development could also be promoted by using engaging interface features (81, 99). Humans tend to use the least amount of cognitive effort when automation or AI-driven decision-support systems are used to support them (107). This means they are more likely to rely on cognitive shortcuts and only superficially attend to the provided information instead of using analytical thinking (108). Decision-makers should be nudged to mindfully attend to the provided information (15) and maintain a certain level of enthusiasm and motivation throughout the decision-making process (109). Interactive aspects can also aid users' ability to experiment with different scenarios of the model outcome and allow a deeper analysis of it (110).

4. Explainability in Expert Contexts

The opaque nature of complex AI models is one of the aspects preventing effective human–AI collaboration (30, 111, 38). Experts are more likely to reject AI suggestions and refuse to adopt DSS if they cannot understand how the system works (88, 30). Experts cannot accurately judge whether they should trust the system (112, 113). They either over-rely on it or choose to systematically disregard algorithmic predictions (71, 87) and follow their old ways of decision-making (36), which are often slower and less accurate than the ones of AI (87, 59). The lack of transparency also makes it complicated to explain how any specific decision was made and whether individuals were treated in a fair, consistent manner and that no biases were introduced (46). In turn, it makes it difficult for an expert to account for their decisions (39).

To solve these issues, AI-generated outputs are often supported with explainability (114, 11, 115). Explainability is a set of technical solutions – more information about explainability techniques and their classifications can be found in Arrieta et al. (11), Došilović et al. (116), and Guidotti et al. (19) – that are intended to provide users with insights into how AI models operate and produce outputs in a comprehensible way (3, 117). Explainability should also reveal the strengths and weaknesses of a decision-making system and enable humans to predict future behaviours (118, 119).

Box 8. Benefits of Explainability in Expert Contexts

- Better understanding of the logic behind the workings of the DSS (120, 121).
- Ability to build meaningful trust and an increased sense of agency (122, 123).
- Reduced cognitive load of performing the task (124).
- Ability to better communicate a final output to the affected parties (e.g., patients) (125, 126).
- Increased willingness to adopt algorithmic systems (127, 128, 129, 130).
- Improved fairness in decision-making (131, 132).

4.1 Explainability in Expert Contexts: Challenges and Limitations

Despite immense research efforts, explainability approaches still lack usability and are ineffective when applied in a decision-making context (133). It has been shown that explainability often does not result in better decision-making – in some instances, it can also lead to undesirable outcomes and can mislead experts (134). Conventional explainability approaches also fail to provide explanations that spark curiosity and motivate experts to learn and solve problems (135). Without adding a clear value to the experts' work, explainability could become a formality and be seen as a redundant feature (136).

Box 9. Explainability Challenges and Limitations

- Explainability can create an illusion of trustworthiness, resulting in experts over-relying on DSSs.
- Explanations not tailored to a specific domain and expert needs will likely be ignored.
- Explanations that require technical knowledge to be evaluated are seen as burdening and frustrating by experts.

4.1.1 Explanations Do Not Fit the Context of the Decision or Are Not Domain-specific

Explanations are often overwhelmingly complex (77) and do not fit within experts' workflows (68). Explainability fails to consider experts' decision-making habits and strategies that they use. For example, Gu et al. showed that introducing DSSs disrupted doctors' ability to make decisions based on historic cases (137). Decision-makers lose interest in explanations when they can't contextualise them or they don't reflect their domain knowledge (99). The standard explainability methods often only provide an 'on the spot' short-term solution but lose their initial value in the long term (138). If explanations are seen as unhelpful and time-consuming, they are more likely to be ignored or inspected superficially (139). Explanations are also likely to be ignored if they are too technical or simplistic (repeat their existing knowledge) (77, 99). Explanations that are simply available but not helpful in promoting understanding feel time-consuming and cause frustration (139).

4.1.2 Understanding Explanations Requires Technical Skills That Are Not Domain Specific

Explainability solutions are often designed with the assumption that experts have a certain level of data science or computational knowledge and skills (140). Thus, many available explainability techniques are too technical for domain experts (141, 142) and are ineffective when introduced in work scenarios (143, 144, 145). For example, research in the medical domain showed that physicians are often unable or unwilling to learn information that is not specific to their domain due to their already intensive workload (137). Explanations are often presented in complex visualisations (145) or numeric representations that require specific skills to be able to interpret them correctly (146). Conejero et al. explored the effectiveness of data visualisation in various governmental decision-making situations (141). The authors used interactive dashboards, charts, maps, and diagrams to illustrate patterns, relationships, and correlations in data. These visualisations were supposed to expose data points a human would not otherwise pick up. However, visual explainability design failed to consider that public administrators in education and employment needed more time or skills to analyse them (141).

4.1.3 Explainability Can Be Misleading

Explanations have been shown to increase blind trust instead of appropriate reliance on AI (138, 147). Explanations can make experts more compliant with the algorithmic systems (99). They might set inaccurate expectations (148) and give an unjustifiable sense of confidence to the decision-makers and make the model seem fairer than it is (149). Explanations that highlight past experiences can result in confirmation bias, which means that experts are more likely to follow the advice that aligns with their opinion rather than the one that challenges it (54). Novice decision-makers have been shown to be especially likely to follow explainable DSS recommendations without challenging them (76, 74, 150, 151). For example, less experienced physicians were more likely to accept incorrect outputs when they were explainable (152). Detailed explanations can strengthen this effect. Comprehensive explanations that included all items from medical history, symptoms, and examination results have been shown to bias primary care practitioners towards AI outputs (77). Moreover, adding explanations can introduce another potential source of error (153).

4.2 Explainability in Expert Contexts: Recommendations

Box 10. Explainability Recommendations

- Explainability should be accessible to domain experts independently of their computational and data science understanding.
- Explainability should be meaningful to experts (e.g., using domain-specific terminology) and fit within their workflows.
- Explainability should be interactive, support flexible information search and outcome comparisons, and provide domain-specific context through examples.
- Techniques, such as cognitive forcing (e.g., asking them to explain their decision), should be used to encourage experts to attend to the explanations and remain engaged.

4.2.1 Ensure That Explainability is Visible

Explainability should be clearly available to domain experts; they should be made aware of how to access and interpret these explanations (154, 155, 156). While experts do not need to know everything – they usually prefer to receive information that is relevant to them – they should be trained to interpret system results and explanations (73). Specific training on explainability could help experts leverage information from these systems more easily. Even informing them about the displays of uncertainty and probabilities would facilitate the increased use of these products (73). Experts can benefit from the additional information that DSSs can provide if it is made accessible and comprehensible to them (88).

4.2.2 Ensure That Explainability is Accessible

Technical explanations do not work in non-technical domains. Explanations should be able to explain the decisions made by the AI in detail to the experts in the field (157). The steps of the AI-driven decision-making process should be accessible and understandable, at least to the expert in the field, who can then explain these rules to the affected individual. This is particularly important in domains where the final decision has a profound effect on the end-user (157). It is essential that these explanations are not only understandable to the data scientists who have the level of mathematical knowledge necessary to understand underlying algorithmic structures. Using explainability methods that show the plausibility of the features used to make a decision does not give the precise reason for the decision. In contrast, the exact decision-making process must be transparent to an expert, for example, a medical professional (157).

4.2.3 Tailor Explanations to a Domain and Task

Explainability should fit with the experts' workflow and use domain-specific terminology that is meaningful to them (18, 16, 143). Linking the terms representing the contributing features to the domain-relevant context (158, 159, 160) and customising explanations to the needs and requirements of experts is effective in making them more usable in supporting mental health practitioners (92, 28) and physicians (99). Explanation in an expert context should reflect what they need to know in each situation and context (158, 159,160). For example, a Tonekaboni et al. study with clinicians using explainable DSS in intensive care units and emergency departments showed that pathologists wanted explanations to show features used to derive the model outcome and areas where the system was most likely to fail (126).

Seeing that allowed experts to compare DSS outputs to their clinical judgement, especially in cases of disagreement (126).

4.2.4 Flexible Information Search Strategies

Explanations should support experts' information search strategies and enable them to find needed information on their use (18). This approach has been shown to be effective in therapy settings (164) and medical contexts (67). Furthermore, explanations that allow flexible information search can help experts find the necessary information in the ways they prefer or are used to (165). In addition, experts feel more motivated and in control when they can freely explore the available explanatory information (166).

4.2.5 Context-specific Explanations

Adding relevant contextual information can help experts to relate to explanations and stimulate their reasoning abilities (167). Furthermore, using design features that prompt users to reflect on their prior knowledge could foster engagement and consolidation of their expert understanding and ability to apply their expertise (168). Including domain-related information to contextualise explanations can improve users' satisfaction and performance (169). Moreover, making explanations more domain-specific can make them more relevant and motivating (99). This approach could also foster experts' ability to recognise similar instances in the future (137). Providing context to AI predictions can also help to interpret DSS outputs more effectively and improve the user's understanding of the model's behaviour (170).

4.2.6 Contrastive Explanations

When decisions are particularly high-risk, explainability should be tailored to challenge experts' fast and intuitive decision-making (171). More analytical decision-making can be promoted by providing explanations comparing different potential outcomes (172). Experts making high-risk decisions tend to match the uncertain situation with their past experiences and evaluate them individually until the necessary information is uncovered (173, 174). This strategy dramatically burdens the decision-maker's working memory (175). To reduce this cognitive load, explainability should show information about the features that influenced AI output sequentially, or by contrasting different outcomes, instead of providing all the explanatory information at once (18). For example, it could deliver explanations that would contrast several options or compare weights of the contributing features. This method helps to highlight distinctive output features and develop an expert ability to notice even salient out-of-the-ordinary events that need to be considered.

4.2.7 Feedback Explanations

Interacting with the system through explanations and feedback has been shown to improve pathologists' ability to engage actively with the outputs of a model and, in turn, provide feature-based feedback that can be used to refine it (164, 176). Explainability that gives feedback on users' performance can help experts to reflect on their own decisions and potentially reduce the potential for bias (138). Interacting with explainability through feedback can create collective, hybrid intelligence on a complex decision-making task with improved accuracy and consistency (164). Moreover, it could encourage experts to mindfully engage with explanations and promote effortful reflection and analysis of the DSS outputs (147).

4.2.8 Collaborate With Experts to Develop Explainability Design Guidelines

It is necessary to uncover user needs or a shared technical understanding. Liao et al., who established an explainability question bank, provided an example where the user's needs for explainability are represented in terms of the questions a user might ask about the AI (84). Another way to understand user needs was proposed by Wolf et al. whereby a scenario-based approach was applied to identify user needs for explainability early in system development (177). Finally, Eiband et al. suggested a stage-based participatory design process, which guides the specification of productspecific needs, i.e., what to explain, followed by the iterative design of solutions, i.e., how to explain (135).

5. Further Questions

There is a need to understand the long-term effects of using DSSs by domain experts. How these technologies will affect expertise development and deskilling can only be extrapolated. However, it remains unknown how experts will adapt to the situation over time and whether they will transfer saved cognitive resources to strengthen the skills not automated by DSSs or develop a new set of expert skills. Understanding how techniques such as explainability could help them do this is essential. This report is an initial step in this direction. However, more efforts must be put into ensuring that experts do not become passive observers of AI work. This is particularly important considering the rapid developments of large language models and generative AI more broadly.

This review revealed that for AI–expert collaboration and explainability to be effective, there is a need for developers and experts to collaborate. A question for future research is how to overcome the initial communication blockages and frustrations and achieve the collaborative stage where both teams align their communication styles. Without directly involving experts in various DSS development and implementation stages, building usable technologies and motivating experts to learn about and adopt them will be challenging. This could lead to wasting resources on technology development and prevent experts from benefiting from DSSs.

References

1. Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., & Wei, L. (2020). *A guide to using artificial intelligence in the public sector.*

2. Procter, R., Tolmie, P., & Rouncefield, M. (2023). Holding AI to Account: Challenges for the Delivery of Trustworthy AI in Healthcare. *ACM Transactions on Computer-Human Interaction*, *30*(2), 1–34.

3. Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, *44*(6), 101976.

4. de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. *Government Information Quarterly*, *36*(4), 101392.

5. Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.

6. Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252.

7. Green, B., & Chen, Y. (2019, January). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the 2019 conference on fairness, accountability, and transparency*, 90–99.

8. Yin, M., Wortman Vaughan, J., & Wallach, H. (2019, May). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.

9. Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational psychiatry*, *11*(1), 108.

10. Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, *102*, 502–520.

11. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R, Chatila. R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, 58,

82–115.

12. Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the social sciences*. arXiv. arXiv:1706.07269 [cs.AI].

13. *Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020, January). Explainable machine learning in deployment.* Proceedings of the 2020 conference on fairness, accountability, and transparency*, 648–657.*

14. Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16).

15. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (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*), 1–21. https://doi.org/10.1145/3449287, accessed 23.05.2024.

16. Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 102792.

17. Croskerry, P. (2003). Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine*, *41*(1), 110-120.

18. Simkute, A., Luger, E., Jones, B., Evans, M., & Jones, R. (2021). Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, 7, 100017.

19. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1–42.

20. Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*, *62*(2), 129–158.

21. Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, *42*(7), 596–615.

22. Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63–71

23. Centre for Data Ethics and Innovation. (2020). *Review into bias in algorithmic decision-making*. https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making, accessed 23.05.2024.

24. Martens, B., & Tolan, S. (2018). *Will this time be different? A review of the literature on the impact of artificial intelligence on employment, incomes and growth*. [JRC Digital Economy Working Paper].

25. Flügge, A. A. (2021, October). Perspectives from practice: Algorithmic decision-making in public employment services. *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, 253–255.

26. McDonald, B., Spaaij, R., & Dukic, D. (2019). Moments of social inclusion: Asylum seekers, football and solidarity. *Sport in Society*, *22*(6), 935–949.

27. *de Sousa, W. G., de Melo, E. R. P., Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda.* Government Information Quarterly*,* 36*(4), 101392.*

28. Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295-305.

29. Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, E., Baltzer, A., Bertlich, I., Blum, A., Bakor-Billmann, T., Bowling, J., ... & Zalaudek, I. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842.

30. Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377.

31. Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce*, *26*(1–2), 147–169.

32. Bertot, J., Estevez, E., & Janowski, T. (2016). Universal and contextualized public services: Digital public service innovation framework. *Government Information Quarterly*, *33*(2), 211–222.

33. Xu, B., Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2019). Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, *22*(6), 1423–1432.

34. Kankanhalli, A., Charalabidis, Y., & Mellouli, S. (2019). IoT and AI for smart government: A research agenda. *Government Information Quarterly*, *36*(2), 304–309.

35. Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10, 113–174.

36. Lee, M. K., Kim, J. T., & Lizarondo, L. (2017). A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-profit organizations. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3365–3376. https://doi.org/10.1145/3025453.3025884, accessed 23.05.2024.

37. Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in human–machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, 89, 27–39.

38. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 671–732.

39. Brauneis, R., & Goodman, E. P. (2018). Algorithmic transparency for the smart city. *Yale J. L. & Tech.*, *20*, 103.

40. *Nair, A. V., Ramanathan, S., Sathiadoss, P., Jajodia, A., & Macdonald, D. B. (2022). Barriers to artificial intelligence implementation in radiology practice: What the radiologist needs to know.* Radiología *(English Edition),* 64*(4), 324–332.*

41. Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022, April). Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–7.

42. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

43. Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*(2), 68–75. https://doi.org/10.1049/iet-its.2016.0208, accessed 23.05.2024.

44. Bolander, T. (2020). What do we lose when machines take the decisions? *Powder Metallurgy and Metal Ceramics*, *23*, 849–867.

45. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access: Practical innovations, open solutions*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052, accessed 23.05.2024.

46. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

47. Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society*, *28*(7), 806–822.

48. Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411, accessed 23.05.2024.

49. Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, *11*(1), 104–122. https://doi.org/10.1002/poi3.198, accessed 23.05.2024.

50. Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2022). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review*, *40*(2), 478–493.

51. Hong, S., & Lee, S. (2018). Adaptive governance, status quo bias, and political competition: Why the sharing economy is welcome in some cities but not in others. *Government Information Quarterly*, *35*(2), 283–290.

52. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, *32*.

53. Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). *Investigating human+ machine complementarity for recidivism predictions*. arXiv. arXiv:1808.09123.

54. *van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations.* Artificial Intelligence*, 291, 103404.*

55. Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in

child welfare services. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300271, accessed 23.05.2024.

56. Wang, X., & Yin, M. (2021, April). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. *Proceeding of the 26th International Conference on Intelligent User Interfaces*, 318-328.

57. Majid, S., Foo, S., Luyt, B., Zhang, X., Theng, Y. L., Chang, Y. K., & Mokhtar, I. A. (2011). Adopting evidence-based practice in clinical decision making: Nurses' perceptions, knowledge, and barriers. *Journal of the Medical Library Association: JMLA*, *99*(3), 229.

58. Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational psychiatry*, *11*(1), 108.

59. Yin, M., Wortman Vaughan, J., & Wallach, H. (2019, May). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.

60. Green, B., & Chen, Y. (2019, January). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the conference on fairness, accountability, and transparency*, 90–99.

61. Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317.

62. Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society*, *2*(2), 2053951715617185.

63. Elwyn, G., Scholl, I., Tietbohl, C., Mann, M., Edwards, A. G., Clay, C., Légaré, F., van der Weijden, T., Lewis, C. L., Wexler, R. M., & Frosch, D. L. (2013). “Many miles to go...”: A systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC Medical informatics and decision making*, *13*(2), 1–10.

64. Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, *21*(4), 70–73.

65. Klein, G. A. (2017). Sources of power: How people make decisions. MIT press.

66. Sterman, J. D., & Sweeney, L. B. (2004). Managing complex dynamic systems: Challenge and opportunity for naturalistic decision-making theory. In H. Montgomery, R. Lipshitz, & B. Brehmer (Eds.), *How professionals make*

decisions. CRC Press.

67. *Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes.* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.* https://doi.org/10.1145/3290605.3300468, accessed 23.05.2024.

68. De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020, April). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

69. Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88–92.

70. Khairat, S., Marc, D., Crosby, W., & Sanousi, A. A. (2018). Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Medical Informatics*, *6*(2), e24. https://doi.org/10.2196/medinform.8912, accessed 23.05.2024.

71. Veale, M., Van Kleek, M., & Binns, R. (2018, April). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14.

72. Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089- 2680.2.2.175, accessed 23.05.2024.

73. Stuart, N. A., Schultz, D. M., & Klein, G. (2007). Maintaining the role of humans in the forecast process: Analyzing the psyche of expert forecasters. *Bulletin of the American Meteorological Society*, *88*(12), 1893–1898.

74. Nourani, M., King, J., & Ragan, E. (2020, October). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 112–121.

75. Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Edridge, F., Mullarkey, D., Buckle, P., & Hanna, G. B. (2021). *Do GPs trust artificial intelligence insights and what could this mean for patient care? A case study on GPs skin cancer diagnosis in the UK*. Preprints. 2021050005.

76. Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019, March). I can do better than your AI: Expertise and explanations. *Proceedings of the 24th*

International Conference on Intelligent User Interfaces, 240–251.

77. Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. *2015 International Conference on Healthcare Informatics,* 160–169. IEEE.

78. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, *4*(1), 31.

79. Howard, A. (2020, March). Are we trusting AI too much? Examining human-robot interactions in the real world. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 1–1.

80. *Borenstein, J., Wagner, A. R., & Howard, A. (2018). Overtrust of pediatric health-care robots: A preliminary survey of parent perspectives.* IEEE Robotics & Automation Magazine*, 25(1), 46–54.*

81. Simkute, A., Surana, A., Luger, E., Evans, M., & Jones, R. (2022, October). XAI for learning: Narrowing down the digital divide between "new" and "old" experts. *Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference*, 1–6.

82. Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88–92.

83. Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63.

84. Liao, Q. V., Pribić, M., Han, J., Miller, S., & Sow, D. (2021). *Question-driven design process for explainable AI user experiences*. arXiv. arXiv:2104.03483

85. Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). *Rethinking explainability as a dialogue: A practitioner's perspective*. arXiv. arXiv:2202.01875.

86. Li, Z., Sharma, P., Lu, X. H., Cheung, J. C., & Reddy, S. (2022). *Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment*. arXiv. arXiv:2204.03025.

87. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them ERR. *Journal of Experimental Psychology: General*, *144*(1), 114. https://doi.org/10.1037/xge0000033, accessed 23.05.2024.

88. Westin, C., Borst, C., & Hilburn, B. (2015). Strategic conformance: Overcoming acceptance issues of decision aiding automation? *IEEE Transactions on Human-Machine Systems*, *46*(1), 41–52.

89. Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300234, accessed 23.05.2024.

90. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504.

91. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kilkin-Gil, R., Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–13. https://doi.org/10.1145/3290605.3300233, accessed 23.05.2024.

92. Yang, L., Wang, H., & Deleris, L. A. (2021, July). What does it mean to explain? A user-centered study on AI explainability. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*, 107–121. Springer International Publishing.

93. *Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., & Van Den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems.* International Journal of Human-Computer Studies*, *154*, 102684.*

94. Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, *20*(7), 1391–1397. https://doi.org/10.1364/josaa.20.001391, accessed 23.05.2024.

95. Balzer, W. K., & Doherty, M. E. (1989). Effects of cognitive feedback on performance. *Psychological bulletin*, *106*(3), 410.

96. Klayman, J. (1988). On the how and why (not) of learning from outcomes. *Advances in psychology*, *54*, 115–162.

97. Klayman, J., & Brown, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, *49*(1–2), 97–122.

98. Hoffman, P. J., Earle, T. C., & Slovic, P. (1981). Multidimensional functional learning (MFL) and some new conceptions of feedback. *Organizational behavior and human performance*, *27*(1), 75–102.

99. Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N., & Ali, R. (2021, October). Nudging through friction: An approach for calibrating trust in explainable AI. *2021 8th International Conference on Behavioral and Social Computing (BESC)*, 1–5. IEEE.

100. Phillips, J. K., & Battaglia, D. A. (2003). Instructional methods for training sensemaking skills. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. National Training Systems Association.

101. Pliske, R. M., Crandall, B., & Klein, G. (2004). Competence in weather forecasting. *Psychological Investigations of Competence in Decision Making*, *40*, 68.

102. Pliske, R. M., McCloskey, M. J., & Klein, G. (2001). Decision skills training: Facilitating learning from experience. In *Linking expertise and naturalistic decision making,* 37-53. Psychology Press.

103. Klein, G. (2008). Naturalistic decision making. Human factors, *50*(3), 456–460.

104. Höferlin, B., Netzel, R., Höferlin, M., Weiskopf, D., & Heidemann, G. (2012, October). Inter-active learning of ad-hoc classifiers for video visual analytics. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 23–32. IEEE.

105. Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C., & Keim, D. A. (2017). What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, 268, 164–175.

106. Dietvorst, B. J. (2016). *People reject (superior) algorithms because they compare them to counter-normative reference points*. Available at SSRN 2881503.

107. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286–297.

108. Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223–241.

109. Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical*

Informatics, *83*(5), 368–375.

110. Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416–428.

111. Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, *316*, 103839.

112. Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

113. Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, 10(1), 85–97. https://doi.org/10.1207/S15327108IJAP1001_5, accessed 23.05.2024.

114. Alicioglu, G., & Sun, B. (2021). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics*, *102*, 502–520.

115. Miller, T. (2017). *Explanation in artificial intelligence: Insights from the social sciences*. arXiv. arXiv:1706.07269.

116. Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. *Proceedings of the 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 0210–0215. DOI:10.23919/MIPRO.2018.8400040.

117. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020, January). Explainable machine learning in deployment. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 648–657.

118. Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850, accessed 23.05.2024.

119. Fuji, M., Nakazawa, K., & Yoshida, H. (2020). "Trustworthy and explainable AI" achieved through knowledge graphs and social implementation. *Fujitsu Scientific & Technical Journal*, *56*(1), 39–45.

120. Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine*, *3*(1), 1–5.

121. VanBerlo, B., Ross, M. A., Rivard, J., & Booker, R. (2021). Interpretable machine learning approaches to prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence*, *102*, 104243.

122. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52.

123. Brennen, A. (2020, April). What do people really want when they say they want "Explainable AI"? We asked 60 stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7.

124. Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European conference on Cognitive ergonomics: The ergonomics of cool interaction*, 1–8. https://doi.org/10.1145/1473018.1473028, accessed 23.05.2024.

125. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain?* arXiv. arXiv:1712.09923.

126. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019, October). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference*, 359–380. PMLR.

127. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

128. Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. https://doi.org/10.1145/2678025.2701399, accessed 23.05.2024.

129. Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118, accessed 23.05.2024.

130. Gilvary, C., Madhukar, N., Elkhader, J., & Elemento, O. (2019). The missing pieces of artificial intelligence in medicine. *Trends in Pharmacological Sciences*, *40*(8), 555–564.

131. DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022, April). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering

harmful algorithmic behavior. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

132. Cheng, C. S., Behzadan, A. H., & Noshadravan, A. (2022). Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Structural Control and Health Monitoring*, *29*(10), e3019.

133. Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2022). *Explainable Artificial Intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival*. OSF Preprints.

134. Jacobs, M., He, J., F. Pradier, M., Lam, B., Ahn, A. C., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., Gajos, K. Z. (2021, May). Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.

135. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018, March). Bringing transparency design into practice. In 23rd international conference on intelligent user interfaces (pp. 211-223).

136. Schemmer, M., Kühl, N., & Satzger, G. (2021). *Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence*. arXiv. arXiv:2109.13827.

137. Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., & Zhang, S. (2020). CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, *40*(2), 699–711.

138. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T, & Weld, D. (2021, May). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

139. van Berkel, N., Skov, M. B., & Kjeldskov, J. (2021). Human–AI interaction: Intermittent, continuous, and proactive. *interactions*, 28(6), 67–71.

140. Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4, 1-25.

141. Conejero, J. M., Preciado, J. C., Fernández-García, A. J., Prieto, A. E., & Rodríguez-Echeverría, R. (2021). Towards the use of data engineering, advanced visualization techniques and association rules to support knowledge discovery for public policies. *Expert Systems with Applications*, *170*, 114509.

142. Woodruff, A., Anderson, Y. A., Armstrong, K. J., Gkiza, M., Jennings, J., Moessner, C., Viegas, F., Wattenberg, M., Webb, L., Wrede, F., & Kelley, P. G. (2020). "A cold, technical decision-maker": Can AI provide explainability, negotiability, and humanity? arXiv. arXiv:2012.00874.

143. Anjomshoae, S., Främling, K., & Najjar, A. (2019). Explanations of black-box model predictions by contextual importance and utility. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1*, 95–109. Springer International Publishing.

144. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv. arXiv:1812.04608.

145. Hadash, S., Willemsen, M. C., Snijders, C., & IJsselsteijn, W. A. (2022, April). Improving understandability of feature contributions in model-agnostic explainable AI tools. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–9.

146. Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*(48), 107539.

147. Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. In C. Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds), *HCI International 2020 – Late Breaking Papers: Multimodality and Intelligence: HCII 2020, vol. 12424*, 449–466. Springer International Publishing.

148. Kocielnik, R., Amershi, S., & Bennett, P. N. (2019, May). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-14.

149. Green, B., & Chen, Y. (2020). Algorithm-in-the-loop decision making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(9), 13663–13664. https://doi.org/10.1609/aaai.v34i09.7115, accessed 23.05.2024.

150. Papenmeier, A., Englebienne, G., & Seifert, C. (2019). *How model accuracy and explanation fidelity influence user trust*. arXiv. arXiv:1907.12652.

151. Wang, X., & Yin, M. (2021, April). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. *26th International Conference on Intelligent User Interfaces*, 318-328.

152. Micocci, M., Borsci, S., Thakerar, V., Walne, S., Manshadi, Y., Edridge, F., Mullarkey, D., Buckle, P., & Hanna, G. B. (2021). Attitudes towards trusting

artificial intelligence insights and factors to prevent the passive adherence of GPs: A pilot study. *Journal of Clinical Medicine*, *10*(14), 3101.

153. Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022, July). How cognitive biases affect XAI-assisted decision-making: A systematic review. *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, 78–91.

154. Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–16.

155. Moehring, F., O'Hara, C. L., & Stucky, C. L. (2016). Bedding material affects mechanical thresholds, heat thresholds, and texture preference. *The Journal of Pain*, *17*(1), 50–64.

156. Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509.

157. Lötsch, J., Kringel, D., & Ultsch, A. (2022). Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics*, *2*(1), 1–17.

158. Lukyanenko, R., Castellanos, A., Samuel, B. M., Tremblay, M. C., & Maass, W. (2021). Research Agenda for Basic Explainable AI. *Proceedings of the 2021 AMCIS conference on information systems,* 1-5.

159. Schaekermann, M., Cai, C. J., Huang, A. E., & Sayres, R. (2020, April). Expert discussions improve comprehension of difficult cases in medical image assessment. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–13.

160. de Greeff, J., de Boer, M. H., Hillerström, F. H., Bomhof, F., Jorritsma, W., & Neerincx, M. A. (2021, March). The FATE System: FAir, Transparent and Explainable Decision Making. *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.*

161. Caro-Martinez, M., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2018). A theoretical model of explanations in recommender systems. *Proceedings of the 2018 ICCBR workshop*, 52-63.

162. Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464–470.

163. Ribera, M., & Lapedriza, A. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. *IUI workshops*, *2327*, 38).

164. Hun Lee, M., Siewiorek, D. P., Smailagic, A., Bernardino, A., & Bermudez i Badia, S. (2023). Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises. *User Modeling and User-Adapted Interaction*, *33*, 545–569.

165. Zehrung, R., Singhal, A., Correll, M., & Battle, L. (2021, May). Vis ex machina: An analysis of trust in human versus algorithmically generated visualization recommendations. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12.

166. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).

167. Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., ... & Jahanshad, N. (2019, March). Towards human-guided machine learning. Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 614-624).

168. Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, *88*(1), 20–32.

169. Bove, C., Aigrain, J., Lesot, M. J., Tijus, C., & Detyniecki, M. (2022, March). Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. *27th International Conference on Intelligent User Interfaces*, 807–819.

170. Baudisch, P., Good, N., Bellotti, V., & Schraedley, P. (2002, April). Keeping things in context: A comparative evaluation of focus plus context screens, overviews, and zooming. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 259-266.

171. Sage, A. P. (1981). Behavioral and organizational considerations in the design of information systems and processes for planning and decision support. *IEEE Transactions on Systems, Man, and Cybernetics*, *11*(9), 640–678.

172. Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition*, *49*(1–2), 37–66.

173. Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14*(5), 331–352.

174. Hutton, R. J., & Klein, G. (1999). Expert decision making. *Systems Engineering: The Journal of The International Council on Systems Engineering*, *2*(1), 32–45.

175. Orasanu, J., & Fischer, U. (1997). Finding decisions in natural environments: The view from the cockpit. *Naturalistic decision making*, 343–357.

176. Roessner, V., Rothe, J., Kohls, G., Schomerus, G., Ehrlich, S., & Beste, C. (2021). Taming the chaos?! Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research. *European Child & Adolescent Psychiatry*, 30, 1143–1146.

177. Wolf, C. T. (2019, March). Explainability scenarios: Towards scenario-based XAI design. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 252–257.

BRAID

Bridging Responsible AI Divides

This fellowship ran from January-May 2023 as part of BRAID.

BRAID is a UK-wide programme dedicated to integrating Arts and Humanities research more fully into the Responsible AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI. Funded by the Arts and Humanities Research Council (AHRC), BRAID represents AHRC's major investment in enabling responsible AI in the UK. The Programme runs from 2022 to 2028. Working in partnership with the Ada Lovelace Institute and BBC, BRAID supports a network of interdisciplinary researchers and partnering organisations through the delivery of funding calls, community building events, and a series of programmed activities. Funding reference: Arts and Humanities Research Council grant number AH/X007146/1.

Learn more at **www.braiduk.org**

This research was supported via UK Research and Innovation by the R&D Science and Analysis Programme at the Department for Culture, Media & Sport. Any primary research, subsequent findings or recommendations do not represent Government views or policy and are produced according to research ethics, quality assurance, and academic independence.

To request an alternative format of this report please email braid@ed.ac.uk.

UKRI Arts and Humanities Research Council

THE UNIVERSITY of EDINBURGH

Ada Lovelace Institute

BBC R&D