THE UNIVERSITY
of EDINBURGH

Reimagining explainability in expert contexts:

Setting foundations for effective expert-AI interactions through design and collaboration

Auste Simkute

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

The University of Edinburgh

2024

Acknowledgements

I would like to express my immense gratitude to the most inspiring supervisors, Professor Ewa Luger, Dr Michael Evans and Dr Rhianne Jones. I am grateful for their continuous support and trust in me, the professional opportunities that allowed me to grow as a researcher, and the freedom to develop my ideas and individual style. Most importantly, I would like to thank my supervisors for being incredibly kind and patient, offering their friendship, and inspiring me by being outstanding researchers. I was fortunate to have the expertise and guidance of my supervisors, who were always happy to help. Lastly, I am incredibly grateful for their support and flexibility during COVID-19, which was very difficult for all.

I am incredibly grateful to the BBC and all the wonderful people working there who welcomed me with open arms and provided me with support and guidance. I want to give special thanks to Dr Bronwyn Jones, who provided valuable advice as an experienced journalist, fellow researcher, and friend. I would like to thank Dr Jones for being my collaborator and helping me write and conduct research. I would also like to thank Professor Adrian Woolard, who welcomed me at the BBC and provided me with endless contacts and opportunities. I would also like to thank Suzanne Clarke for helping me run the workshops with UX designers across the BBC and for generously supporting and assisting me throughout the process. I would also like to thank all the workshop participants who found time in their busy schedules to participate and engage with the tasks enthusiastically. I am also grateful to everyone at the Rosalind Franklin Institute who kindly welcomed me, shared their workspace with me, participated in my study, and generously shared their knowledge.

Lastly, I am thankful for my loved ones. First, I thank my wonderful and loving parents, who provided me with the support I needed to live and study in Scotland and were always encouraging, even when my ideas seemed foreign to them. I am also profoundly grateful for my wonderful friend, Tyler, who helped me during the most challenging times of my PhD. I would also like to thank my best friend, Monika, who has always been my biggest supporter. I am grateful for my friends Fil, Anna, Aditi and Kasia, who patiently listened to all my ideas and always uplifted me.

Abstract

Domain experts increasingly rely on Artificial Intelligence (AI) systems to support their decision-making. However, they often struggle to build meaningful trust in opaque and complex technologies or find them poorly fitting within their workflows. As a result, experts either over-rely on or refuse AI systems instead of benefiting from them. Explainable AI (XAI) has been seen as one of the solutions to support experts' interactions with AI, but despite extensive research efforts, XAI techniques have not been effective in practice. This thesis argues that expert-orientated explainability should stem from an ongoing collaboration and mutual understanding between experts and stakeholders involved in its development. Explainability should also empower long-lasting learning and experts' ability to use their expertise through design informed by empirical knowledge from Human Factors and Cognitive Psychology research.

I first review XAI development in recent years and how its focus shifted from developing predominantly technical to more user-centred solutions, from targeting narrow groups of technical users to a broad range of stakeholders with varying levels of technical backgrounds. Then, based on an ethnographically informed study with experts and software developers, I outline foundations for explainability set through collaboration and feedback. Then, I draw on the Human Factors research literature review and discuss how this knowledge could be used to build systems that would empower experts. Informed by the reviewed literature and ideation workshops with interface designers, I present a conceptual design framework to align explainability interface features with expert decision-making strategies in varying risk and time-pressure contexts. Based on Cognitive Psychology and Human Factors research literature, I introduce learning and cognitive engagement strategies for explainability that would motivate experts and foster expertise development. Finally, I connect all three parts, showing how all these aspects are necessary for explainability to be effective in expert contexts.

Lay summary

Domain experts, such as physicians and social workers, increasingly rely on Artificial Intelligence (AI) systems at their jobs. For example, a medical expert might use AI to screen biopsy results and find atypical instances, and a social worker might rely on AI to screen unemployment cases and flag the cases at risk of long-term unemployment. However, AI systems are too complex and opaque to be inspected by a human. As a result, domain experts often misuse AI technologies or refuse to use them. Using techniques designed to explain AI systems and their outputs is seen as one of the ways to improve user understanding and AI adoption. However, these techniques are rarely effective.

In my thesis, I first explore why domain experts cannot use AI systems successfully and why they do not benefit from the provided explanations. To find answers, I reviewed literature from various research disciplines, including Human-Computer Interaction, Cognitive Psychology, and Human Factors Engineering. The reviews revealed that introducing AI systems disrupts experts' ability to use their expertise and often does not fit within their workflows. Explainability solutions fail to support expertise and are rejected as too technical or irrelevant to the task or a specific domain. Moreover, current explainability solutions can lead to overreliance and biased decision-making.

Then, I address the underlying issues unrelated to system design that could lead to ineffective explainability, such as experts' inability to use the AI system. Based on an ethnographically informed study with expert scientists and AI developers, I outline recommendations to increase AI adoption. I argue that domain experts need to receive appropriate support when starting to use new technologies, receive adequate training, and continuously learn about AI systems by collaborating with AI developers. These steps should form a foundation for the introduction of explainability.

Next, I address how explainability could support experts and help them integrate their expert knowledge with the AI system's suggestions. In response to this issue, I developed a conceptual design framework that should inform user interface designers working on explainable AI interfaces. This framework links expert-preferred decision-making strategies and contextual factors, such as time and risk, with the design suggestions that are most suitable for each situation. I then provide a

list of concrete design suggestions generated during workshops with the user interface designers.

The literature reviews also revealed that experts find explanations frustrating and unhelpful and, as a result, do not engage with them. They also showed that a knowledge gap is forming between new experts who have a high level of AI literacy but low domain expertise and old experts who have low computational knowledge but high domain expertise. To address these challenges, I provide design recommendations based on Cognitive Psychology techniques that promote learning and engagement. The recommendations are intended to increase the XAI potential of teaching computational skills and provide opportunities to develop expert skills.

Overall, in my thesis I argue for a holistic approach to explainability and provides design guidelines and recommendations for making explanations enabling, engaging, educative, and based on collaboration between experts and AI developers.

Table of contents

Chapter 1 Introduction

1.1 Background

Aspirations for more accurate and robust predictions led to the widespread applications of complex Artificial Intelligence (AI) techniques, such as random forests and deep neural nets (Adadi & Berrada, 2018). At the same time, AI systems have become more available than ever before (Grace et al., 2018), as many public and private sector organisations have been drawn to AI's capabilities to efficiently gather, process, and analyse data, as well as perform various complex tasks more effectively than if done manually (Vedapradha et al., 2019). As a result, AI is now ubiquitous in expert domains such as finance (Cao et al., 2021), the public sector (Reis et al., 2019), journalism (Diakopoulos, 2019), natural science research (Baum et al., 2021), the healthcare sector (Hamet & Tremblay, 2017), policing (Hung & Yen, 2021), the legal sector (Brooks et al., 2020) and education (Zhang & Aslan, 2021).

1.1.1 Issues with opaque AI systems in expert contexts

However, advancing AI technologies are increasingly more complex and incomprehensible to humans. Inherently interpretable but arguably less accurate models are often traded for opaque, black-box models that can be unintelligible even to AI engineers and data scientists, let alone regulators, auditors, users, and individuals affected by AI decisions (Barredo Arrieta et al., 2020). As a result, it has become increasingly challenging to ensure that AI decisions are justifiable, unbiased, and fair (Goodman & Flaxman, 2017) and that humans operating these systems have the means to notice and correct potential errors, including instances of algorithmic unfairness (Zhao et al., 2017). Moreover, users who need AI support to perform their tasks are often reluctant to rely on technologies that are incomprehensible to them due to their opaqueness or high complexity, even when it means that their work progress is stifled (Yu et al., 2017). On the other hand, they might over-rely on AI outputs if they cannot evaluate them accurately (Green & Chen, 2019a).

1.1.2 Expansion of XAI research field

Given the critical contexts in which AI systems are embedded, the need to make them understandable for various audiences snowballed, attracting much attention

from Human-Computer Interaction (HCI), Computer Science (CS) and AI research communities (Abdul et al., 2018). Interest in developing eXplainable AI (XAI) solutions further grew after pressure from legislators and, in particular, Europe's General Data Protection Regulation (GDPR), which introduced a "Right to Explanation" to ensure that those affected by algorithmic decision-making have a right to question algorithmic decisions (Goodman & Flaxman, 2017). This meant that algorithmic decision-making had to maintain a meaningful human input that could then justify its decisions to those affected by them (Selbst & Powles, 2017). In response to the emerging demand, multiple technical solutions have been proposed to make black-box algorithms explainable to their users (Guidotti et al., 2019). Researchers also started to recognise that XAI is a multidisciplinary issue and began extrapolating from other disciplines, such as Social Sciences, to build an understanding of how people explain and perceive explanations in everyday situations (Miller, 2019) and how these preferences might differ in the context of interacting with AI (Weisz et al., 2019).

1.1.3 Explainability in expert contexts

Despite the immense research efforts, XAI solutions have been shown to lack usability when implemented in real-life situations (Abdul et al., 2018), especially in professional settings where AI is used to support domain expert (e.g., physicians) decision-making (Bhatt et al., 2020). Although XAI has been seen as one of the solutions to support domain experts' interactions with AI, it became apparent that this stakeholder group has more complex explainability needs than anticipated. For example, domain experts not only have to make judgements about AI output correctness but also to integrate AI-provided information with their expert knowledge. The ability to use their expertise often depends on their workflows, contextual factors and the information they receive (Klein, 2008). All these factors are affected when AI is introduced.

Moreover, domain experts might not have the high AI or data science literacy necessary to use AI systems and interpret their outputs effectively. As a result, they often find generic explanations that are not tailored to their needs either overwhelming or too simplistic (Bussone et al., 2015; Naiseh et al., 2021). Introducing explainability in expert domains without ensuring that it is effectively used by experts can have undesirable results and lead to overreliance on AI. For example,

explanations can create an illusion of transparency and make an AI system seem fairer than it is (Ananny & Crawford, 2018) or nudge users to rely on AI outputs instead of their expertise (Buçinca et al., 2021b). However, little has been done to understand how explainability could more effectively support domain experts in interpreting AI outputs and integrating them with their knowledge. Research has also not sufficiently explored how explainability could improve experts' ability to interact with AI systems and build meaningful trust in them and their outputs.

1.1.4 Identifying research gaps

In 2019 and 2020, the XAI field was predominantly focused on supporting data scientists and software developers. There was little research evidence on how domain experts interact with AI tools and how explainability could have been used to support them. During the scoping review of XAI research literature (Chapter 2), I recognised that solutions proposed to domain experts often failed to acknowledge the fact that domain experts not only needed to understand AI outputs but also had to integrate them with their own knowledge. To my knowledge, no research at the time specifically focused on XAI's impact on experts' ability to apply and develop their expertise when making decisions with the support of AI. Furthermore, literature reviews also revealed that domain experts experience difficulties effectively adopting and using AI tools, meaning that explainability features cannot be used by them as they are not able to use the AI system in the first place. However, HCI, AI, and XAI researchers tackled these two issues separately without considering that successful AI adoption should be a prerequisite to explainability, as without it, explanations will simply not be needed.

1.2 Thesis motivation, aims and objectives

The overarching motivation for this thesis is to address the issue of explainability in expert domains from a wider perspective without focusing on a specific technical or theoretical solution. Instead, the motivation is to inform the XAI research field about the broad spectrum of aspects that influence the effectiveness of XAI in expert contexts, and that should be considered when designing XAI methods or interfaces for domain experts. I define this broader approach as a **holistic perspective**.

1.2.1 The holistic approach to explainability

To reiterate, the holistic perspective means that explainability solutions should encompass multiple parts of expert-AI interaction that influence whether and how experts use explainability and whether and how they benefit from it. I argue that for explainability to be effective in expert contexts, researchers should consider the whole process of expert-AI interaction and think broader than the explanation itself. In this thesis, I outline the three core aspects of the holistic approach that contribute to XAI's effectiveness in expert contexts:

- Explainability should have a strong foundation and stem from an ongoing collaboration between domain experts and other stakeholders (e.g., software developers). Domain experts should be made aware of system capabilities and limitations. They should also receive adequate training and support when starting to use a new system and continue exchanging knowledge with developers and other assigned team members. This part is covered in Chapter 4.
- Explainability should be designed to enable experts to exercise their expertise. The explainability interface design should be tailored to support experts' sensemaking and decision-making strategies and reflect contextual aspects, such as time pressure and risk. This part is covered in Chapter 5.
- Explainability should be meaningful and engaging. This means that explanations should be designed to capture domain experts' attention and increase their willingness to engage with the provided information. Explanations should also provide valuable information in a way that promotes long-term knowledge development. Chapter 6 covers this part of the holistic approach.

The holistic approach was chosen due to the complexity of domain expertise. This group of stakeholders has a wide range of skills and knowledge, preferred workflows, and might work under challenging circumstances (e.g., high time pressure or risk). Domain experts might also have either high or low AI or data science literacy, making it difficult to predict their ability to interpret explanations. As previously mentioned, experts also have to integrate their knowledge with AI outputs while also making judgements of their correctness. Lastly, people working in expert domains need to develop and maintain their expertise by practising their tasks. AI

automation without explainability can stifle their ability to do so. I argue that these complexities specific to expert contexts require an expansion of XAI beyond just explaining AI outputs and the underlying algorithmic processes. With this thesis, I wish to provide an informative knowledge base outlining the complexities of expert-AI interactions and specificities of expert explainability needs for those who are designing, developing and researching XAI to domain experts.

1.2.2 The research aims and objective

In Chapter 2, I aim to develop an overview of the XAI technical perspective and provide an overview of the role domain experts play in the CS and AI research literature on XAI in comparison to other stakeholders. I argue that CS and AI research can help to understand the state-of-the-art XAI methods and the research questions guiding their development, which in turn can help to understand how they support domain experts and what the key limitations are. In this chapter, I also aim to recognise directions for the subsequent stages of the literature review and answer the question of what the gaps in the XAI research literature are regarding explainability for domain experts.

The key objective of Chapter 2 is to identify the limitations of current XAI methods and techniques in supporting domain experts and develop the research agenda for increasing XAI usability and effectiveness in expert contexts.

In Chapter 3, I aim to explore the key motivations for explainability in expert domains and develop a clear understanding of the role XAI plays in ensuring algorithmic fairness, accountability and effective transparency in the context of AI. Moreover, in this chapter, I aim to understand the changing expert work dynamics and barriers that prevent them from effectively using AI systems. I argue that it is important to take a step back and deepen our understanding of the challenges arising from human-AI interactions in expert contexts before proposing new solutions. I encourage researchers to learn from the domain of Human Factors Engineering (HFE), which has been exploring human-automation interaction issues for over two decades. I aim to extrapolate from HFE research to understand better how automation impacts domain expert workflows and the ability to use their expertise. For example, HFE offers extensive empirical knowledge on aspects such as trust-related biases and experts' adaptation strategies to new technologies being introduced to support them at work. More precisely, I aim to develop a knowledge

base explaining the trust dynamic in human-AI interaction and the role of explainability in supporting meaningful trust, as well as to develop a knowledge base helping to understand the challenges and possibilities of expert-AI interactions and the role of explainability in supporting experts' ability to stay in the decision-making loop.

Chapter 3's objective is to provide a knowledge base on expert-AI interaction challenges, such as problematic trust dynamics and difficulties staying in the decision-making loop, based on a multidisciplinary literature review that also extrapolates from decades of HFE research literature. This knowledge base will inform AI, HCI, and XAI researchers and others aiming to effectively support expert-AI interactions.

In Chapter 4, I aim to draw attention to the underlying issues of low AI adoption among domain experts. I argue that before introducing solutions such as explainability, we need to ensure that experts are able to access and effectively use AI systems. Based on an ethnographically informed study with expert scientists and software developers, I aim to i) Investigate the practical and contextual barriers that experts face when new AI systems are introduced into their workflows, the challenges to successful AI use and adoption occurring in different stages of AI system introduction, including before experts start using them, ii) outline the perceptions of the reasons for low in-house AI software adoption among experts, AI developers, and team leaders within the science organisation, iii) identify how different stakeholders' perceptions align or misalign and how that influences the adoption of an AI system, iv) understand how AI developers, team leaders and experts approach the issue of low AI adoption; v) explore the role of collaboration between stakeholders for the AI adoption.

The key objectives of this chapter are to propose solutions to support effective AI adoption based on the results of the ethnographically informed study and outline the aspects of expert-AI interaction that need to be addressed before XAI methods are introduced into their workflows.

In Chapter 5, I propose explainability-focused design solutions. In the first part of the chapter, I aim to outline how explainability could be designed to address these challenges and more effectively support domain experts by identifying expert decision-making strategies and factors that influence them. I aim to provide a

conceptual framework to inform and show how this framework could be applied in practice in domains that are less researched, such as journalism.

In the second part of Chapter 5, I aim to improve the ERT framework by recognising aspects that might be challenging for designers and aiding the understanding of how it could be made more usable for design practitioners and decision-makers. I also aim to match proposed design goals with concrete design approaches gathered during the workshops with UX practitioners, populate the ERT framework with examples of design strategies, and shape informative design guidelines for explainability interface design.

One of the objectives of Chapter 5 is to provide a list of design suggestions for XAI that would improve expert-AI interactions and inform UX designers, XAI developers, and AI researchers. Another objective is to develop a conceptual framework that would help tailor explainability interfaces to support experts' information needs under various contextual circumstances and accommodate workflow disruptions due to the introduction of AI. Lastly, I wish to provide a list of concrete design suggestions and practices developed based on the outcomes of the design workshops with UX practitioners. This list will help designers and other stakeholders to inform their practices when designing XAI interfaces.

In Chapter 6, I aim to propose methods that would make explainability more than just a tool helping domain experts decide whether to rely on certain outputs. I argue that explainability can be tailored to augment decision-makers' abilities and help them benefit from the available technologies. I aim to outline the temporality and engagement issues related to explainability in expert contexts. I also aim to inform explainability researchers and designers about the methods they can apply to increase the educational potential of explanations and make them more engaging and sustainable. I argue that by reflecting on learning methods from the Cognitive Psychology research field, we can build explanations that would support expertise development and preserve valuable expert skills in the context of AI.

The key objectives of this chapter are to provide a list of design suggestions that would make explanations engaging and educative and to provide a framework illustrating when engaging explanations should be shown to domain experts.

1.2.3 Disclosure of changes to research plans due to COVID-19

At the beginning of my PhD studentship, the expectation was to conduct more empirical research with journalists at the BBC. In particular, I was supposed to stay at the BBC for six months, working with journalists and news editors, conducting ethnographic research and collaborating with people working within the organisation. However, ethnographic research was improbable because during and after COVID-19, experts worked mainly from home, which was seen as a temporary solution at the time and was not considered an ecologically valid study environment. The study plans were first delayed and then cancelled. I then aimed to conduct remote interviews with journalists and editors at the BBC. Still, due to their increased workload, my interview requests were continuously rejected, or scheduled interviews were cancelled at the last minute. As a result, my main focus became literature reviews and the development of the conceptual framework and design guidelines. First, I collaborated with a researcher and experienced journalist, Dr Bronwyn Jones, who helped me embed my framework into the context of journalism. Then, I conducted workshops remotely with UX designers at the BBC and design students at the University of Edinburgh.

Another aspect that complicated studying AI applications in journalism was the way it was used at that moment in time. Researchers at the BBC were developing AI tools for journalists and editors, but these tools were still prototypes and were not used at work. The way journalists used existing data-driven tools was minimal and difficult to study in an observational manner. For example, a journalist might use AI to summarise or translate information and explainability was not seen as part of the process. However, given the new developments of AI tools, it was clear that explainability will be important in the future, and it will have to be developed to fit the busy and fast-paced workflows and routines of journalists and editors. Also, their ability to judge the AI-provided information is highly important to ensure news quality, information integrity, and the reputation of the BBC. Thus, the development of the conceptual work and informative guidelines using speculating scenarios and knowledge from the UX designers working at the BBC still seemed very valuable.

After conducting these studies, I aimed to explore another aspect of the holistic approach to explainability. I was hoping to understand why experts who need AI tools often either do not use them or do not use them effectively. The large natural

science organisation was recognised as the potential case study during the public presentation of its AI tools. Their representative mentioned that scientists at the organisation do not use the in-house tools that they need to process the amount of data they now collect. The organisation fit the criteria for the case study because i) AI systems were essential for experts' progress (as the volumes of data are too large to process manually), but the adoption of the AI-enabled software remains low, ii) the advances in data collection technologies were still recent, and experts were new in having to rely on AI systems, iii) practitioners and experts worked in the same building, and some projects required their collaboration. The organisation invited me to conduct the study and observe real-life challenges domain experts face when AI tools are introduced in their workflows, as well as the role explainability plays in the process.

1.2.4 Thesis overview

Chapter 1 (this chapter) presents the background and key motivations for the research and defines the aims and objectives of the thesis.

Chapter 2 introduces the technical XAI background. It reviews XAI developments in recent years and illustrates how the focus of explainability researchers shifted from developing predominantly technical solutions for data scientists and developers to a broad range of stakeholders with varying levels of technical backgrounds. This chapter also demonstrates the issues with applying technical approaches to address challenges involving people who might not have technical expertise. There are a number of advanced XAI methods available, but they lack a connecting link (e.g., tailored interface design) that would make them usable to the domain experts.

Chapter 3 draws from multiple disciplines, such as HCI and HFE and introduces the first in-depth and multifaceted literature review on the challenges decision-makers face when trying to stay in the loop. This chapter builds an understanding of why it is important that experts are supported and can be a meaningful part of AI-driven decision-making. It also discusses issues related to trust in AI and analyses underlying reasons for automation bias and algorithmic aversion. Lastly, it discusses issues related to changes in experts' roles and how introducing AI can disrupt their working patterns and workflows.

In **Chapter 4**, I introduce an idea of foundations for effective explainability. I report on the ethnographically informed contextual inquiry study with experts and software

developers. Based on the study results, I outline foundations for explainability set through collaboration and feedback. I define four AI acceptance requirements that could help to overcome the obstacles to effective AI adoption and present recommendations for effective collaboration to support implementations of AI systems in workplaces.

Chapter 5 builds on the previous two literature reviews (Chapters 2 and 3) and discusses how explainability interfaces could be designed to enable expertise in AI-supported decision-making. Based on the reviewed literature, I present a conceptual design framework to align explainability interface features with expert decision-making strategies in varying risk and time-pressure contexts. I demonstrate how this framework could be applied in practice by using speculative scenarios informed by the journalism domain. Lastly, I populate this framework with design suggestions generated during the ideation workshops with UX designers.

In **Chapter 6**, I introduce an idea of extended XAI value. I argue that explanations are often “shallow” and only provide on-the-spot solutions but do not contribute to long-term knowledge development. Without seeing a clear value in explanations, domain experts are unlikely to engage with them meaningfully. Drawing from Cognitive Psychology and HFE research literature, I propose learning and cognitive engagement strategies to inform explainability design. I argue that engaging and learning-focused explanations could help narrow the divide between “new” and “old” experts by promoting technical and domain knowledge development.

In **Chapter 7**, I discuss the findings and connect the thesis's outcomes to the aims and objectives raised in this Chapter. I also outline this research's key contributions and propose future research directions.

Chapter 2 Explainability background literature: Overview of XAI research trajectory and critical technical solutions

Chapter 2 summarises the background literature on explainability and introduces the key explainability concepts, techniques, and approaches. This chapter covers i) the terminology of the explainability used across the research publications, ii) the critical aspects of technical explainability, iii) the most popular XAI techniques and types of explanations, iv) the explainability stakeholders, and v) the key challenges for explainability in expert domains. This chapter is based on the literature review of explainable AI research published during 2015 – 2020, when the topic of explainable AI gained initial attention from various research communities. This part of the thesis shapes the direction of the further chapters of this thesis and recognises the lack of research attention for XAI in expert domains. Chapter 2 also aims to help the readers understand the technical potential of explainability and see the underlying reasons why XAI techniques lack usability in expert domains.

2.1 Chapter introduction

Rapid AI advances and wide application of its opaque techniques
urged researchers and developers to work on finding the best ways to explain how complex systems, such as random forests and deep neural nets, work while maintaining their high learning performance (Adadi & Berrada, 2018). This led to the development of various algorithmic techniques designed to explain black-box models or their outputs using post hoc explanations (Barredo Arrieta et al., 2020). The post hoc techniques can be applied to explain the unique output of a model or the model itself (Guidotti et al., 2018). Some cutting-edge approaches involve learning a simple local approximation of the underlying models around a particular data point (Ribeiro et al., 2016), producing an additive feature importance score for single predictions (Lundberg & Lee, 2017), identifying how a given prediction would need to change for an alternative outcome to occur (Mothilal et al., 2020), or removing a particular instance from the training data would impact the decision boundary (Bueff et al., 2022). These methods are usually used in combinations and complement each other to deliver the most complete and accurate explanations (Belle & Papantonis, 2021). In general, explainability is intended to make the basis behind a system's reasoning in making a prediction comprehensible to humans (Office, 2020). It should also

reveal the strengths and weaknesses of a decision-making system and enable humans to predict its future behaviours (Gunning & Aha, 2019).

There are various ways explanations can be displayed to the user. For example, XAI techniques can be expressed in natural language, symbols (Weber et al., 2018), visualisations (Tamagnini et al., 2017), and by using examples or extracting representative instances from the training dataset to demonstrate how the model operates (Cai et al., 2019). Researchers also recognise that different stakeholders have varying needs for explainability (Ribeiro et al., 2018). Researchers recognise the need to tailor explainability to different stakeholders. However, it is unclear how XAI methods could be effectively implemented in non-technical domains (e.g., social work). Most technical XAI methods require a certain understanding of data science to be interpreted and lack usability in domains where users have little or no computational knowledge (Bhatt et al., 2020).

2.2 Methodology

The literature reviews for this thesis were conducted in three separate stages (this chapter represents the literature review stage one) while gradually gaining an understanding of the state of the XAI research and recognising gaps in the research literature. The gradual approach was used given that at the time of conducting the first stage of the review, XAI studies were predominantly focused on technical methods despite it being a highly multidisciplinary issue. The research regarding domain experts' explainability needs was almost non-existent. Thus, my intention was to expand my literature review beyond the domain of CS and AI or the XAI research.

2.2.1 Stage 1 – Review of the XAI research literature

CS and AI research can help to understand the state-of-the-art XAI methods and the research questions guiding their development. However, conducting a system literature review of this research was not in line with the aims of this thesis, as it is not intended to develop new technical methods or focus on evaluations or algorithmic improvements of the existing ones. Moreover, several large-scale and in-depth systematic reviews covering XAI techniques have already been published. Thus, the first stage of the review is informative but not systematic.

The literature review for this stage was guided by the keywords recognised throughout the key XAI reviews. The keywords ‘intelligibility’, ‘explainability’, and ‘interpretability’ (with a condition of AI or ML also appearing in the abstract), as well as ‘intelligible AI/ML’, ‘explainable AI/ML’, and ‘intelligible AI/ML’ were used for the search.

For this stage, the research literature was found using the University of Edinburgh digital library DiscoverED, ACM, IEEE Xplore, The Science Direct and JSTOR databases. Literature search during the first stage was limited to the publications made in 2015-2020. This was done to reflect on the most up-to-date literature at the time the review was conducted and with a consideration that the interest in this research area rapidly grew within these four years.

The key aims and research questions of this part of the literature review are:

1. **Provide an overview of the XAI technical perspective to understand the role domain experts play in the CS and AI research literature on explainability**
    a. What are the key definitions used to refer to explainable AI? What is the state-of-the-art definition?
    b. What are the key concepts used in XAI research?
    c. What are the main types of explainability?
    d. What are the main methods of explainability?
2. **Recognise the stakeholders in XAI research literature and the role of domain experts**
    a. What are the key stakeholders mentioned in the XAI research literature?
    b. In what context domain experts are referred to in the XAI research literature? How is their role described in the XAI context?
    c. What are the differences between XAI methods suggested to domain experts and other stakeholders?
    d. How do existing XAI methods support domain experts?
3. **Recognise directions for the subsequent stages of the literature review**
    a. What are the gaps in the XAI research literature regarding explainability for domain experts?

2.3 XAI terminology classification

Complex and wide-spanning needs for explainability meant that the field quickly became highly multidisciplinary and polarised (Abdul et al., 2018). This diversity and rapid growth have resulted in a lack of consistency across publications, especially regarding the terminology used by stakeholders such as researchers, media, and software developers (Lipton, 2018). The growing accessibility of AI tools and their overwhelming prevalence in non-technical domains increased the diversity of its potential users, further complicating the field of XAI (Tomsett et al., 2018). Multiple definitions were used, sometimes referring to the same and sometimes to different concepts across publications, preventing continuity in research and instead highly scattering the field (Poursabzi-Sangdeh et al., 2021). The consensus on terminology was seen as necessary to promote unity and scientific rigour in the field, which has started becoming highly multidisciplinary (Abdul et al., 2018; Doshi-Velez & Kim, 2017; Lipton, 2018). This section aims to explain the critical definitions of *intelligibility*, *interpretability*, and *explainability* that have been most prevalent across research and other communication channels (e.g., media and legislation) and also highlight the division and lack of communication within the multidisciplinary XAI community.

2.3.1 The terminology in AI and other research communities

The initial terminology confusion was followed by the surge of publications trying to untangle multiple concepts and definitions often used interchangeably (Clinciu & Hastie, 2019). However, the terminology studies often did not follow each other. They only introduced more complications instead of establishing a consensus until some prominent publications set a tone by using specific definitions in their highly cited and influential works (Barredo Arrieta et al., 2020; Gunning & Aha, 2019). The terminology confusion made it difficult for disciplines to collaborate and effectively communicate, further dividing their approaches to the problem. For example, technical research papers used terms such as interpretability and explainability to reference the specific technical properties of AI models (Zhang et al., 2018). On the other hand, the HCI community and disciplines, such as law or journalism, used the same terminology more generally to define any explanation of AI output or a system (in undefined form) to the intended stakeholders (Selbst & Barocas, 2018). However,

over time, even technical publications (solely focusing on algorithmic solutions) started resorting to a single definition for simplicity (Belle & Papantonis, 2021; Bueff et al., 2022).

2.3.1.1 Intelligibility

In the context of AI systems, this term refers to “the way of communicating a complex computational process to a human” (Weld & Bansal, 2018) or the system’s ability “to expose the inner workings and inputs of context-aware applications that tend to be opaque to users” (Lim et al., 2019). The term *intelligibility* is sometimes used aside from similar definitions (Abdul et al., 2018) or as a broader term incorporating qualities of *interpretability* and *explainability* (Weld & Bansal, 2018). The latter terms are more specific, and although they have often been used interchangeably and considered to be closely related (Adadi & Berrada, 2018) or even synonymous (Dhurandhar et al., 2018), they are also argued to have distinct meanings (Gilpin et al., 2018).

2.3.1.2 Interpretability

In the context of AI systems, *interpretability* constitutes “the ability to explain or to present in understandable terms to a human” (Doshi-Velez & Kim, 2017). More precisely, *interpretability* is used when the aim is to provide a comprehensible description of the internals of a system (Gilpin et al., 2018). This term is often used in the AI- and XAI-techniques-oriented research literature (Doshi-Velez & Kim, 2017). It is also more often used among AI scientific communities than in public settings (Adadi & Berrada, 2018). These communities use *interpretability* to define a passive characteristic of a model that is comprehensible to a user (assuming they have the necessary technical knowledge), corresponding to the model’s algorithmic transparency (Ribeiro et al., 2016; Rudin, 2019). *Interpretability* can also be used to define methods and models that display and make clear the behaviour and predictions of AI models (Belle & Papantonis, 2021), which means that the underlying workings of the models are visible and inspectable to a user to make sense of the reasoning behind the AI decision. *Interpretability* is usually used when the focus is on explaining a model and its inner processes with an interest in domain experts rather than end-users (Doshi-Velez & Kim, 2017).

2.3.1.3 Explainability

In the context of AI systems, the term *explainability* indicates the intelligent system's ability to explain its strengths and weaknesses, support human comprehension of the system's workings and outputs, and predict its future behaviours (Gunning & Aha, 2019). In the research literature, the aim of *explainability* is defined as "to ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to end-users and other stakeholders in non-technical terms" (Selbst & Barocas, 2018). *Explainability* is mostly used with a focus on human understanding of algorithmic outputs through effective explanations (Gunning & Aha, 2019). This term is also more often used in public settings (Adadi & Berrada, 2018) and official guidelines (Burt et al., 2018). *Explainability* usually defines a more user-centred approach that seeks to deliver comprehensible information to the end-users about how the AI outputs inform algorithmic decisions (Selbst & Barocas, 2018). When used within a technical context, *explainability* refers to an active characteristic of a model, meaning that the model takes deliberate actions to explain its reasoning to be understandable to a human (Chakraborti et al., 2020). Currently, the research community is settled on the term eXplainable AI (XAI), which defines an area of research concerned with tools and techniques that make AI systems and their outputs understandable to their users, building trust between humans and black-box models (Gunning & Aha, 2019).

2.4 XAI technical perspective

This section briefly covers the technical side of the XAI research field (algorithmic XAI techniques). The technical XAI methods have been extensively reviewed and surveyed in recent years (Adadi & Berrada, 2018; Guidotti et al., 2019; Gunning & Aha, 2019; Murdoch et al., 2019). However, as the algorithmic side of explainability is not the main focus of the thesis, it will not be explored in depth here but will provide an overview needed to establish an understanding of the basics of technical XAI capabilities, available models, and their key goals. This section is intended to help a reader understand the technical resources available that can be complemented by various design and user interface approaches and the suggestions and guidelines provided in this thesis.

2.4.1 Opaque and transparent AI models

In the XAI context, AI models are divided into two essential categories and are either accounted as *transparent* or opaque *black-box* models. These categories refer to the level of complexity of the model and whether it is inspectable without applying additional techniques and is comprehensible to the human user (i.e., the model is transparent) (Du et al., 2020). They might also refer to whether the algorithmic information of the model is made available to its user (i.e., the model is transparent) (Barredo Arrieta et al., 2020). For the AI model to be transparent, its procedures of generating outputs should be visible and comprehensible without applying additional techniques (Ahmad et al., 2018). In other words, the transparent model's algorithmic information is available to the user to inspect through a mathematical analysis, so they could put different input points and understand how the model arrived at the output. Black-box models that fail to meet transparency criteria (e.g., neural networks) need an additional method to be applied to it in order to interpret its reasoning. These additional methods are XAI techniques intended to communicate the underlying workings of the model clearly and understandably. Because these techniques are applied to already trained models, they are called post-hoc techniques (Mittelstadt et al., 2019).

2.4.2 Key XAI techniques

Broadly, the XAI techniques can be either model-agnostic and model-specific. Model-agnostic techniques can be applied to any AI model, whereas model-specific techniques are designed for a particular model or family of models and thus are restricted to them. More precisely, a range of XAI techniques are available to either explain various AI models or to be applied to specific ones (Barredo Arrieta et al., 2020; Gunning & Aha, 2019).

2.4.2.1 Shapley Additive exPlanations (SHAP)

Shapley Additive explanations (SHAP) are the most popular XAI technique for explaining individual predictions (Lundberg & Lee, 2017). It is a model-agnostic value estimation method that uses ideas from the game theory to measure feature attributions. It aims to build a linear model around the instance being explained and then interpret the coefficients as the feature's importance. The SHAP explanations are usually expressed as visualisations that the user needs to have specific

statistical and data science knowledge to inspect and interpret. This technique has been criticised as some models might be “immune” to these explanations. The SHAP technique has also failed to uncover biases (Slack et al., 2020).

2.4.2.2 Counterfactuals

Counterfactual explanations are provided as statements identifying how a given prediction would need to change for an alternative outcome to occur (Mothilal et al., 2020). These explanations show the minimal change variables required to predict the alternative outcome. This can be presented as a similar case, for example, another successful job application, to compare against the one that AI rejected. It helps the user to understand what led to one application being accepted and the other being rejected by the model.

2.4.2.3 InTrees

InTrees is a model-agnostic XAI technique for tree ensembles (Deng et al., 2017). This method uses the tree architecture to produce explanations and functions as a pipeline of multiple algorithms extracting, measuring, pruning, selecting, and summarising rules. The key idea behind this technique is that although a tree ensemble might be opaque, each of its constituents is transparent and can be inspected.

2.4.2.4 Deletion diagnostics

Deletion diagnostics is the XAI technique that investigates the model as a function of its training data (Bueff et al., 2022). It considers how removing a particular instance from the training data would impact the decision boundary. This technique can reveal the influential instances in the training set, so it is beneficial for model debugging, fixing mislabelled data, reducing the training set, and generally understanding what features the model relies on for predictions.

2.4.2.5 LIME and Anchors

Lastly, *LIME and Anchors* (Ribeiro et al., 2016, 2018) are classifiers that are considered easy to inspect and comprehend. LIME explains the predictions of any classifier by learning an interpretable model locally around the prediction (Ribeiro et al., 2016). It can present text as a binary vector indicating the presence or absence of a word. For an image explanation, it can produce a binary vector indicating the “presence” or “absence” of a contiguous patch of similar pixels (a super-pixel). At the

same time, the classifier may represent the image as a tensor with three colour channels per pixel. The Anchor's technique is based on the principle of simple if-then rules to describe a model's reasoning. They explain individual predictions locally by identifying a decision rule that "anchors" the prediction. A constraint anchoring a prediction implies that changes to the remaining feature values do not impact the prediction. Anchors require a perturbation-based strategy and the black-box model to generate local explanations. This technique is believed to be accessible to non-technical stakeholders due to its simple propositional rules and visualisations (Belle & Papantonis, 2021).

An agreement in the AI community is that no single technique can answer all the questions and provide a complete picture of model reasoning, but they can complement one another. Thus, it is argued that these techniques should be combined when appropriate (Belle & Papantonis, 2021). However, one needs special training and an understanding of when and how each technique can be useful to achieve that. It is also a question of whether a user would have access to all these techniques and how time-consuming, expensive and complicated it would be to apply multiple techniques to understand an instance. Moreover, a non-technical audience might be able to understand Anchor's technique but would struggle with interpreting more complex visualisations of SHAP. Lastly, an important note should be made that these explanations must be accurately interpreted to be useful, requiring a certain level of data science and statistical knowledge.

2.4.3 Types of explanations

Explanations produced by applying XAI techniques can be used to either explain a unique output of a model (i.e., local explanation) or the model itself (i.e., global explanation). In some cases, explanations fall in between the two categories.

Local explanations aim to provide a clear explanation for a trained model by explaining a specific input of it (Guidotti et al., 2018). These types of explanations do not intend to include all the available information about the model but are limited to explaining a single input. Local explanations can be model-specific if they are designed to explain a specific instance in a deep learning model or model-agnostic if they generate model-agnostic explanations for a particular instance or the vicinity of a specific instance (Rai, 2020).

On the other hand, *global explanations* are intended to explain a global model behaviour and require information about the trained model, such as the algorithm and data used to train it. Global explanations are based on a general perspective of the input features and the learned coefficients, such as weights, parameters, and network architecture. These explanations make the importance of model features and their interactions comprehensible to the user and provide clarity about the distribution of the target feature conditioned on the input features (Lundberg et al., 2020). Global explanations can also be either model-specific or model-agnostic. Model-specific explanations are achieved by incorporating interpretability constraints into the structure of the model, for example, by using fewer features as inputs or by constraining the relationship between features and predictions to monotonic (Rai, 2020). Model-agnostic explanations use an approximation of an interpretable model for the black-box model, e.g., a deep learning model can be appropriated to the decision tree (Rai, 2020).

The global and local explanations can then be presented to the user in various ways. For example, explanations can be expressed using natural language, rules, and symbols (Weber et al., 2018). Explanations can also be visual and might rely on data visualisation techniques, such as clusters, colour graphs, and saliency maps (Tamagnini et al., 2017). Visualisations are used with the intention to direct the user's attention to the decision boundary, feature interactions, or other essential aspects of the model (Tamagnini et al., 2017). Explanations can be delivered using examples, extracting representative instances from the training dataset to demonstrate how the model operates (Cai et al., 2019). Explanations can also be based on the feature relevance. For example, features might be ranked using importance scores, thus giving the user insight into which features were essential to arriving at the particular model's behaviour (Barredo Arrieta et al., 2020).

2.4.4 XAI stakeholders

The diversity of potential AI users who need explanations has encouraged researchers to look for ways to categorise XAI stakeholders and their needs. Researchers started recognising that a user-centred approach is needed to support various stakeholders and provide relevant information to them. The wave of interest in XAI began with the research focus being on data scientists and software engineers, as primary stakeholders (Belle & Papantonis, 2021). However, the

increase in the model applications in sensitive decisions and mounting evidence of the risk of using black-box models for these decisions expanded the stakeholder scope.

Ribeiro et al. (Ribeiro et al., 2018) grouped potential stakeholders into three main areas based on users' goals, backgrounds, and relationships with the product. They suggested that XAI should be tailored to users based on their role in the AI ecosystem. They argued that developers and AI researchers, e.g., software developers, should receive explanations to help verify the system, detect failures, and improve it. The explanations should apply to an audience that can understand the code, data representation structures, and statistical deviations. Domain experts, e.g., physicians, should receive explanations using natural language or interactive visualisations, allowing the explainability to be guided by the user. Lay users, e.g., patients who received an AI-informed diagnosis, should be provided an explanation in the form of counterfactuals (Ribeiro et al., 2018).

In another prominent XAI review, Barredo Arrieta et al. (2020) divided target audiences in XAI into five key groups needing explainability for different purposes. The authors expanded some of the categories proposed by Ribeiro (Ribeiro et al., 2018) and added regulatory entities and managers. Their proposed stakeholders were: i) domain experts and users of the model (e.g., physicians) who need XAI to build trust in the model and gain scientific knowledge, ii) regulatory entities and agencies that need XAI to certify model compliance with the legislation and audits, iii) managers and executive board members who need to assess regulatory compliance and understand corporate AI applications, iv) users affected by the AI-driven decisions who need to understand how AI is affecting their situation and to evaluate if the decision was fair, lastly v) XAI is needed by data scientists, developers, product owners to improve the model continuously and to improve product efficiency and conduct research. Similar reasoning was proposed by (Adadi & Berrada, 2018), who suggested, that looking at the purposes of the XAI is needed. The authors did not specify XAI categories, but they suggested that different explanations should be used: i) to justify and ensure that AI decisions are not erroneous, ii) to improve models or control them by preventing things from going wrong, and iii) to help users learn about the system and discover new things so that they can learn new aspects of hidden laws in biology, chemistry and physics (Adadi & Berrada, 2018).

A different approach was suggested by Ehsan et al. (2020), who introduced a user-centred XAI approach by applying a sociotechnical method, which incorporates both technical and social elements needed to understand the potential user. They conducted a case study using an online game and asking participants to think aloud. The authors observed that user perceptions differed depending on their professional and educational backgrounds and the relationships within teams' collaborative decision-making. Overall, they proposed that XAI should be designed depending on the social factors surrounding AI systems. They also suggested that a holistic understanding of the users is needed, considering their values, interpersonal dynamics, and the socially situated nature of AI systems. Wang et al. (2019) also suggested user-centric explainable AI design and looked at the cognitive biases and explanations suitable to mitigating them. However, the authors did not consider contextual influences and how these cognitive biases and explanation requirements might differ across stakeholders, domains, and contexts.

2.5 Chapter overview

XAI topic has attracted a lot of attention across domains, and the number of research publications in 2017-2019 has grown rapidly (Abdul et al., 2018; Barredo Arrieta et al., 2020). The considerable interest and the pace at which this growth happened led to confusing and interchangeable definitions and concepts dividing the AI technical community from other disciplines, such as Law and HCI. Due to the rapid publication pace, XAI research rarely built on each other's findings, which led to tangled terminology and confusion about what information is essential to what stakeholders, when and how explainability should be applied, and what the critical reasons for explainability are (Adadi & Berrada, 2018). Many technical XAI solutions have been proposed and successfully applied in data science and software development communities. However, most of the XAI techniques are designed with the thought of a data scientist as the key stakeholder, rarely considering other potential contexts and specific needs of diverse stakeholders (Abdul et al., 2018). XAI methods often include complex visualisations that require a certain level of statistical and data science knowledge to interpret them accurately. This suggests that for individuals without this experience, XAI techniques could only add an extra layer of complexity and increase the potential of an error. Moreover, in expert domains, explainability is

often seen as an additional feature that requires effort and time to interpret and lacks usability (Bhatt et al., 2020).

The field of XAI recognises that explanations should differ depending on the user who receives them. However, besides the study by Ehsan et al. (2020), researchers mainly suggested extensive stakeholder categories with few potential purposes. There has been little attention to what stakeholders want to know, what they can understand given their expertise, or what contextual constraints might prevent them from evaluating explanations or engaging with them. Regarding domain experts as a stakeholder category, both Barredo Arrieta et al. (2020) and Ribeiro et al. (2018) only provide a very brief rationale for the broad needs of experts, such as building trust and gaining scientific knowledge (Ribeiro et al., 2018). Few studies focused on the specific explainability needs of experts, and even then, the researchers only looked at aspects such as recommendations for using a domain-relevant vocabulary (Ribeiro et al., 2018). Little attention has been paid to understanding the environments in which these systems are implemented and observing how experts interact with AI at work. Some early attempts have been made to draw insights from social sciences on how humans explain and understand explanations (Miller, 2019; Wang et al., 2019), and few studies recognised the need for user-centred approaches (Ehsan & Riedl, 2020; Wang et al., 2019), but little has been done to explore how people (especially domain experts) reason in partnership with algorithms and what factors might interfere with their natural sensemaking processes. Moreover, little consideration has been given to designing and effectively implementing XAI to enable or preserve expert knowledge. In the following chapters, I will explore these knowledge gaps and focus on reasons for XAI and why some approaches might not be effective for experts' needs at the cognitive and contextual levels, as well as aspects of engaging experts' attention and enabling them to integrate their expertise and AI suggestions.

Chapter 3 XAI opportunities and challenges in supporting human-in-the-loop

Chapter 3 explores the key motivations for effective explainability in expert domains and outlines the aspects that challenge an expert ability to stay in the loop. This chapter covers: i) the importance of meaningful human agency in AI-supported decision-making, ii) the role experts play in enabling responsible AI applications, iii) the expert trust dynamics and trust-related biases, iv) the challenges of experts staying in the loop, v) and how explainability can help to overcome these challenges. This chapter is grounded in a literature review exploring research publications on human-in-the-loop, AI in decision-making contexts and XAI in decision-making contexts. It is also further supported by the literature review of the HFE research publications on automation in expert domains. This part of the review provides a deeper understanding of the potential challenges of humans staying in the loop. This chapter grounds the reasoning of the following chapters, as it shows why it is essential to support domain experts by helping them understand how AI works and ensuring they can continue applying their expertise when using intelligent systems. This chapter also shapes the research questions of Chapters 5 and 6, as it analyses the human-in-the-loop challenges and shows that experts' cognitive and contextual needs should guide effective explainability.

3.1 Chapter introduction

AI has become increasingly prevalent in domains and settings where humans are expected to have a certain level of expertise. AI-driven technologies are used to inform expert judgements in domains such as policing (Brown et al., 2019), the healthcare sector (Gaube et al., 2021; Procter et al., 2023), recruitment (Flügge, 2021), governmental work (Duan et al., 2019; Wirtz et al., 2019), and social services (Chouldechova et al., 2018), to name a few. For example, in healthcare, AI can be used to support cellular pathologists in making diagnostic decisions based on biopsy results. In this case, the AI system would be used to screen out typical results and inform physicians about the atypical cases (Procter et al., 2023). In social services, AI can be used to profile an unemployed individual's case by calculating their risk score based on variables like education, age, gender, and type of housing. The AI system would analyse these scores and determine the programme the applicant is

eligible for (e.g., job placement, vocational training, apprenticeship, activation allowance). It would warn a social worker to inspect if someone is at risk of long-term unemployment (Flügge, 2021; Martens & Tolan, 2018). In immigration services, AI can be used to screen individual applications for a visa or refugee status and recommend to the immigration worker whether a person should be allowed to enter a country and which cases should be evaluated with additional scrutiny (Kuziemski & Misuraca, 2020; McDonald & Spaaij, 2021).

These highly sensitive decisions require expert ability to oversee and challenge AI effectively. If a decision-maker cannot understand how AI produces certain outputs or their decision-making is disrupted, there is a risk that experts will be pushed out of the decision-making loop (De-Arteaga et al., 2020; Elwyn et al., 2013; Janssen et al., 2019a). The sensitivity of AI-driven decisions in domains such as healthcare or police also poses questions about ensuring algorithmic fairness (Chouldechova et al., 2018; Yang & Stoyanovich, 2017). AI outputs might be biased and erroneous, and if these biases are not noticed, inspected and challenged, they can escalate and have detrimental and discriminatory effects at individual and societal levels (Crawford & Schultz, 2014). This raises questions about the transparency of AI-driven decisions and whether an expert can know if the data is biased when the information about the model and its output is incomprehensible (Diakopoulos & Koliska, 2017).

Moreover, if the expert in the loop cannot understand the workings of the AI or does not have the time or means to inspect the AI output, it is unclear who should be held accountable in case of an error (Binns, 2018). Without being able to understand how the system works, experts are unable to make informed decisions about whether to trust the AI and its outputs or not (Bussone et al., 2015; Yu et al., 2017) and often perform worse with the support of AI than on their own (Amann et al., 2020; Gaube et al., 2021; Micocci et al., 2021). Misuse of AI systems can lead to inaccurate decisions that can be particularly costly in high-risk and sensitive domains (Kankanhalli et al., 2019; Zhang et al., 2020). Without meaningful human input, algorithmic unfairness might remain unrecognised until the targeted investigation is conducted (Mattu, 2022) and might lead to the replication and even amplification of existing biases in society (Bolander, 2019; Zhao et al., 2017).

Explainability could play a significant role in ensuring that experts have the means to interact with AI and maintain their agency and expertise. It could help decision-

makers understand the logic behind AI predictions and enable meaningful agency by providing relevant information (Cutillo et al., 2020; VanBerlo, 2021). Explainability can also help them understand how a system works, which is necessary for trusting a newly introduced AI system (Brennen, 2020). However, research shows that explainability instead can lead to overreliance (Bansal et al., 2021).

3.1.1 Publications

This Chapter is based on the report *Explainability in Expert Contexts: Challenges and Limitations in Supporting Domain Experts in AI-driven Decision-making*. I researched and wrote the report as part of the Department for Culture, Media and Sport Junior Policy Fellowship, and it is in the process of being published. The subsection regarding automation bias is based on the short opinion paper *Explainability and Trust-Related Biases,* which was accepted and presented at the ECSCW'22 workshop *Building Appropriate Trust in Human-AI Interactions*.
The subsection on the challenges of staying in the loop is informed by the HFE literature review used for the publication *Ironies of Generative AI: Understanding and mitigating productivity loss in human-AI interactions*, which is currently undergoing peer review.

3.2 Methodology

The first stage of the Literature Review (Chapter 2) revealed that different stakeholder needs, especially those of domain experts without the technical knowledge necessary to interpret XAI explanations, were not explicitly considered. This suggested that there is a need for a better understanding of their specific explainability needs and the rationale of why effective explainability is important in expert contexts. However, due to the lack of expert-focused research in the explainability area, I expanded my literature review beyond XAI to explore motivations for explainability in expert contexts and, more generally, to understand domain experts' requirements for technologies such as AI. To understand the expert-AI and expert-XAI interaction challenges when new systems are introduced into their workflows, I also expanded the literature search to include the discipline of HFE.

3.2.1 Review of the research relevant to the XAI in expert contexts (Stage 2)

During the second stage of the review, I first identified parallel research themes relevant to expert contexts (e.g., human-in-the-loop, meaningful trust) by analysing the most influential publications in the XAI research field. The key publications were recognised during the first stage of the literature review (e.g., (Abdul et al., 2018; Barredo Arrieta et al., 2020; Guidotti et al., 2019; Lipton, 2018; Miller, 2023). They were the most cited at the time of the review and did not focus only on the technical aspects of the XAI. These reviews included analyses of the most important research directions in the field of XAI, from which I identified relevant research themes that were not directly about XAI but were helping to answer the research questions of this review and either explored aspects relevant to expert-AI interaction or were motivations for XAI to be applied in expert contexts.

The key motivations mentioned in the research literature were 'accountability', 'fairness', 'transparency', 'human-in-the-loop' and 'trust'. They were used as keywords and paired with 'AI', 'ML', and 'algorithmic decision-making' during the search. The additional keyword 'decision-maker' was selected as the majority of the expert domain concerning research literature, referring to AI in decision-making contexts instead of AI in expert contexts. Literature search during the first stage was limited to the publications made in 2015-2022. The Boolean logic was applied to the following electronic databases: Web of Science and Scopus. Google Scholar was used to manually search for further relevant, influential, and most cited publications. The final sample of empirical studies was selected through a stepwise process. The list of papers was manually filtered, excluding publications that did not investigate the expert/decision-maker as a stakeholder or did not align with this review stage's aims and research questions.

The key aims and research questions of this part of the literature review are:

1. **To understand the role of explainability for experts in ensuring algorithmic fairness, accountability and effective transparency in the context of AI**
    a. How can explainability help to ensure algorithmic fairness in expert contexts?
    b. How can explainability help to ensure accountability in expert contexts?

c. How can explainability help to ensure effective transparency in expert contexts?

2. To understand the trust dynamic in human-AI interaction and the role of explainability in supporting meaningful trust

a. How do experts build trust in AI outputs, and how can explainability support them and help to develop meaningful trust?
b. What are the effective and ineffective aspects of explainability in expert contexts?
c. To understand the challenges and possibilities of expert-AI interactions and the role of explainability in helping experts stay in the decision-making loop
d. What are the research gaps and potential future strategies of decision-making systems and explainability in expert contexts?
e. How much agency experts have in AI-informed decision-making? What role does explainability play in enabling this agency?
f. What are the key challenges of maintaining a meaningful expert role?
g. How can explainability support the human-in-the-loop approach?

3.2.2 Subsequent review of the Human Factors literature

3.2.2.1 Trust dynamics in expert contexts

During this review stage, the important and (at the time of the review) underexplored research areas became apparent. The research exploring human-AI interaction in the second stage of the review revealed the importance of meaningful trust and humans' ability to judge the trustworthiness of AI and its outputs. However, the research was still premature in the area, with only a few publications covering the issues of automation bias, algorithmic aversion, or exploring trust dynamics between humans and AI. At that time, to my knowledge, no publications explored these questions regarding domain experts or conducted them in the real-world context, observing domain experts' trust dynamics in the context of AI. Mitigating automation bias and avoiding algorithmic aversion appeared to be important motivations for explainability. However, the aspects of XAI that could nudge experts to overtrust the AI system also surfaced.

To fill this gap, I explored HFE research publications, as this field of research has studied human-automation interaction for almost three decades. In particular, HFE

has extensively focused on expertise in various decision-making contexts. It has studied domain experts (e.g., aircraft pilots, air traffic controllers, plant operators) and their interactions with automation, including how the introduction of automation affects experts' decision-making, workflows, ability to exercise their expertise and strategies that could be used to support experts' ability to maintain agency and effectively interact with automation (Parasuraman et al., 1993; Parasuraman & Riley, 1997). HFE also researched biases related to the introduction of automation, such as automation bias. The latter has also been occurring in AI contexts (Gaube et al., 2021).

First, I explored the key publications on trust and automation. Based on the terminology used in these papers, I conducted the search using the keywords "automation bias", "complacency", and "overreliance". I used Google Scholar and DiscoverED, the University of Edinburgh digital library, to conduct the search. I included publications that were transferable to the context of human-AI interaction, i.e., considered higher-level dynamics between humans and automation, that were not specific to a particular interface or task. Some of the HFE research literature referred to very practical aspects, such as firefighters having to react to unforeseen circumstances, or it considered specific interfaces, such as dashboard details in an aeroplane cockpit. These publications were excluded. I included the most cited publications that considered trust dynamics between experts and automation and answered one or more of the following questions:

- *What leads to automation bias?*
- *How do trust dynamics change over time?*
- *What strategies can help to alleviate automation bias/build meaningful trust?*

3.2.2.2 Human-in-the-loop and expert-automation interaction

Another aspect that was revealed as an important motivating factor for explainability in expert contexts was the human ability to stay in the loop. Reviewed HCI, AI, and XAI research literature identified that humans are not always able to remain in the decision-making loop, covered why it is important that humans stay in the loop, what it means, and how explainability could, in theory, help achieve it. Still, it did not cover why users, especially domain experts, struggle to stay in the loop or explore the underlying challenges in depth. As a response, I explored the key HFE publications

on experts integrating their knowledge and using their expertise when automation is introduced into their workflows. This subsequent review of HFE research literature aimed to inform the research space and better understand why experts find it complicated to exercise their expertise and maintain meaningful control when AI tools are introduced into their workflows and effective explainability requirements in expert contexts.

As with the automation bias literature, I recognised the terminology used in the seminal papers on human-in-the-loop and conducted the search using keywords of “human-in-the-loop” and “automation” combined with “decision-making”. The search was conducted using Google Scholar and DiscoverED, the University of Edinburgh electronic library. Similar to the previous search, I included publications that were transferable to the context of human-AI interaction, i.e., considered higher-level dynamics between humans and automation, that were not specific to a particular interface or task. The publications that focused predominantly on practical aspects or specific interfaces rather than the interaction with the technology were excluded. I included publications that considered the human ability to use their expertise when automation was introduced, including effects on their workflows and decision-making strategies. I focused on the most cited (seminal) publications that provided both conceptual/theoretical and empirical understanding of the challenges in human-automation interaction that affect humans' ability to stay in the loop.
This part of the review was guided by the following questions:

- *What interferes with an expert’s ability to stay in the loop?*
- *What factors affect their ability to integrate their knowledge with the information the automation provides?*
- *What helps experts stay in the decision-making loop?*

3.3 Importance of a meaningful human agency/human-in-the-loop

A human-in-the-loop agent can supervise and control the automated process (Rahwan, 2018). Rahwan (2018) named two main functions human agents perform in AI-driven decision-making. The first is to identify misbehaviour by an otherwise autonomous system and take corrective action when needed. The second is to provide an accountable entity in case the system misbehaves. To ensure that

humans have the necessary agency to perform these two functions, it is essential to support experts interacting with AI systems (Raghu et al., 2019).

Sensitive decisions, such as who will receive refugee status in a country, are highly complex and discretionary and require human oversight and input (Janssen et al., 2022; Kuziemski & Misuraca, 2020). Humans are essential in making these decisions as they can show empathy, consider unusual circumstances, and notice salient factors not reflected in the training data (Bolander, 2019; Procter et al., 2023). AI is rarely intended to replace humans in making these decisions. Instead, the fundamental goal of AI is to augment the decision-making process instead of fully automating it (Hong & Lee, 2018). Ideally, AI would automate mundane, repetitive tasks and allow experts to focus on higher-level and creative ones (Zanzotto, 2019). However, humans should stay in control of making the final decision, while automation is used to aid it. Having a human overseeing the workings of AI has been shown to reduce errors in healthcare (Raghu et al., 2019) and legal sectors (Tan et al., 2018). Experts themselves also express a wish to maintain control and autonomy in decision-making (Brown et al., 2019; van der Waa et al., 2021). Helping experts stay in the loop is a way to ensure that their expertise and knowledge are considered and that sensitive decisions are not left to full automation.

The human-in-the-loop approach is partly driven by the growing societal appeal for responsible AI (Barocas & Selbst, 2016). Besides the potential to improve the accuracy of these decisions, algorithms are often linked to social, ethical, and legal issues, and even unfairness towards certain groups or individuals (Barocas & Selbst, 2016), as well as the lack of accountability (Diakopoulos, 2015) and transparency (Ananny & Crawford, 2018). AI-driven tools are susceptible to errors due to biased or incomplete datasets (Janssen & Kuk, 2016; Selbst & Barocas, 2018). AI systems can also be insensitive to the context and might ignore essential factors that should be accounted for when generating the output (Bolander, 2019). AI-driven systems are also trained to find statistical correlations. Such correlations might or might not be a result of a causal relationship. For example, AI might suggest a meaningful connection even if no causal association exists (Bolander, 2019). Having humans in the loop can be an essential safeguard for responsible AI applications. Affected members of society also prefer maintaining human agency in AI-driven decision-making – they believe that humans can better ensure consideration of any unusual or salient factors when making decisions (Brown et al., 2019; Lee et al., 2017).

To stay in the loop, experts need to be able to build meaningful trust in AI. Meaningful trust can be described as the ability to make an informed judgement about whether or not to rely on an AI output based on sufficient information and the ability to analyse and understand it. When experts cannot form meaningful trust in an output that is not explained to them or if the explanation is incomprehensible, their trust is driven by heuristics. For example, decision-makers might decide whether to rely on AI depending on how the model was introduced to them (Alon-Barkat & Busuioc, 2023; Lee & See, 2004), how it performed during the first few trials (Sauer et al., 2016), how experienced the decision-maker is (Gaube et al., 2021), and whether the output confirms their expert intuition (Ghassemi et al., 2021).

3.3.1 Algorithmic fairness

Fairness in the context of AI-driven decision-making concerns algorithmic practices being implemented in a way that would ensure just and non-discriminatory outcomes (Yang & Stoyanovich, 2017). The key goal is to prevent AI from directly and indirectly affecting individuals and groups (and society) and preventing them from being treated unequally based on sensitive characteristics, such as gender, race, or age (Binns et al., 2017). However, due to AI-driven technologies' opaque and complex nature, unfairness can go unnoticed and cause significant harm (Burrell, 2016; Selbst & Barocas, 2018). Algorithmic unfairness can result from biases in the historical data (Barocas & Selbst, 2016; Zhao et al., 2017), appearing due to the small size of the training dataset (European Parliament. Directorate General for Parliamentary Research Services., 2019) and various institutional practices, the background, and the culture of the data subjects whose data are gathered (Janssen et al., 2022).

Allowing AI systems to function without human oversight can lead to discriminatory effects and the perpetuation of societal biases. For example, ProPublica’s COMPAS risk assessment system analysis demonstrated how it discriminated against black defendants (Mattu, 2022). After analysing data from more than 10,000 criminal defendants, authors found that black defendants were more likely to be falsely assessed as high risk and less likely to be incorrectly assessed as low risk to re-offend than white defendants (Mattu, 2022). In another example, gender inequality was propagated by historically biased algorithms when Google’s job search engine showed men ads for jobs with higher pay than women (Datta et al., 2015), even in

“lower stakes” domains (e.g., media or entertainment), the consequences of algorithmic errors can be costly and influence the quality of life and well-being at both individual and societal levels. For example, in journalism, the use of news recommender systems has prompted concerns about their role in limiting access to diverse content by creating filter bubbles and echo chambers that may be detrimental to democracy and polarise societies (Helberger et al., 2020).

3.3.1.1 Algorithmic fairness through XAI

Effective explanations could be used to design user-driven algorithm audits and help leverage users’ prior experience of harm and biases (DeVos et al., 2022a). For example, when XAI was used to enable decision-makers to make holistic risk assessments and adjust for algorithmic limitations, it reduced the racial disparity in child welfare screening compared to when AI was used alone (Cheng et al., 2022). XAI could also help to recognise unfair practices and provide feedback about them. For example, van Berkel et al. (2019) assessed how users perceived fairness using an explainable recidivism risk prediction system. Participants in the study demonstrated the ability to interpret underlying data from graphs generated using an XAI technique and to make an informed decision about the fairness of the predictions made by the system. Participants of the study were also able to provide informed feedback about the system. Having a human-in-the-loop via XAI practices has also been shown to effectively reduce error rates in medicine (Raghu et al., 2019) and increase fairness in recidivism risk predictions (Tan et al., 2018).

On the other hand, explainability can be misleading and distract users from fairness issues. A qualitative study by Binns et al. (2018) revealed that users’ opinions on the fairness of algorithmic decisions can change depending on the type of explanation provided to them. When study participants were presented with different types of textual explanations of algorithmic decisions performed in high-level legal decision-making contexts, they differed in their judgements. When participants were repeatedly exposed to a single type of explanation, these differences disappeared as people presumably paid more attention to the case rather than aspects of explanations. Thus, exposing the user to multiple explanations might distract them from the fairness issue and disrupt their reasoning process.

In a similar study, Dodge et al. (2019) showed that explaining the values of different features and providing similar examples from the data set might more

effectively reveal fairness discrepancies between other cases. In contrast, the explanations that show the decision boundaries or the structure of the training data and how it is distributed concerning the decision boundary could be used to enhance the general perception of fairness of the model. Authors suggested that combining both types could lead to the most balanced evaluation of algorithmic fairness. Moreover, findings by Dodge et al. (2019) also indicated the role of individual differences and their prior attitudes towards the AI systems' fairness in how explanations impacted their judgements. This effect was present when participants showed general bias towards algorithm fairness and when they expressed preference for a particular feature used. These results suggest a need for more user-centric and personalised approaches in designing explainability interfaces and explaining algorithms to users when questioning the system's fairness.

3.3.2 Algorithmic transparency

Algorithmic transparency is the accessibility of the model's inner workings, which can be inspected using mathematical techniques. Transparency is also understood in a broader sense as the level to which the information about the system's reasoning and data management practices is made available to the public (Ananny & Crawford, 2018). Making the AI system visible by revealing the algorithmic processes behind it could be a way to introduce more accountability and fairness into decision-making (de Laat, 2018). Transparency is helpful to regulatory bodies, whose role is to oversee and decide whether these systems should be deployed in the first place. However, when it comes to the individual decision-makers, it can be overwhelming (de Laat, 2018). Access to the algorithms and datasets can also increase the risk of revealing private information, damaging the company's competitiveness and leading to system manipulation or overwhelming decision subjects (de Laat, 2018).

3.3.2.1 Transparency without explainability is not enough

Transparency does not, by default, allow the user to interact dynamically with the system and build a deeper understanding of its workings (Ananny & Crawford, 2018). There is also a risk that inadequate transparency attempts would create a transparency illusion, i.e., enable a false perception of an open and fair system whilst insufficient to deliver accountability and prevent algorithmic discrimination. These performative efforts to be transparent might give an organisation unjustified

credibility (Ananny & Crawford, 2018). As a solution, Annany and Crawford (2018) suggested that algorithmic systems should not only be made open and visible but also comprehensible. For example, the authors proposed that allowing users to explore the system could help them understand how algorithms produced a specific output and, in turn, challenge those algorithms (Ananny & Crawford, 2018).

Providing transparency by simply revealing algorithmic information to the user might not necessarily give them agency and means to make sense of the available data. For example, transparency cannot guarantee that an intended stakeholder will be able to assess the system's accountability and fairness if the visible information is not tailored to their information needs (Kemper & Kolkman, 2019). Similarly, transparency cannot help build meaningful users' trust in the system without revealing information relevant to the specific user (Schnackenberg & Tomlinson, 2016). Goodman and Flaxman (2017) criticised the original idea of open algorithms. Instead, they proposed that making relevant information available and understandable to intended audiences would be more effective. These findings point to the potential and need for tailored explainability approaches to make transparency practical and accessible to various stakeholders, such as domain experts. This could be achieved by making relevant information available and understandable to the user. For example, it has been suggested that users' expectations for transparency should be evaluated, and based on the result, the relevant information should be selected (Brown et al., 2019).

3.3.3 Accountability

Accountability in AI-driven decision-making refers to the human decision-maker being able to provide justifications for the design and operations of the algorithmic system used for making a decision (Binns, 2018). However, in some cases, it is unclear whether decision-makers or designers, developers of algorithms, or their management should be held accountable for potential consequences. Diakopolous (2015) argued that due to the importance of the human role embedded into algorithms during the design and interpretation stages, algorithmic accountability should fall on the parties involved in both stages. However, decisions are often left unaccounted for (Moses & Chan, 2018). For example, in predictive policing, where algorithms are used to guide the deployment of police resources to high-crime-risk locations, algorithms changed the accountability of individual officers and law

enforcement agencies. The agencies now often struggle to provide a substantial explanation of who should be held accountable in the case of an accident (Bennett Moses & Chan, 2018). On the other hand, when the regulations put a responsibility on decision-makers, they might not have sufficient means to understand and override the decisions made by the AI system (Wagner, 2019). In areas such as autonomous driving or social media, moderators were often held responsible for the outcomes even when they had little agency to interfere with or override the decisions made by algorithms (Wagner, 2019).

3.3.3.1 XAI's role in ensuring accountability

Explainability could introduce more meaningful accountability into AI-driven decision-making and allow experts to evaluate AI suggestions and make an informed judgement about whether to rely on them or not to be justly held accountable. Diakopolous (2015) analysed five case studies in journalism and reported challenges related to applying accountability methods in practices. In the study, one recognised challenge was the journalists' lack of technical skills necessary to understand and evaluate the AI-driven system. Diakopolous (2015) suggested that to ensure accountability, some specific parts of the model, such as the features and variables and their weights used in these algorithms, should be made available and explainable to the users or parties inspecting these algorithms. Similarly, Rahwan (2018) discussed several methods for how transparency mechanisms can achieve algorithmic accountability. One of the methods they recognised was making information about the system behaviour and interactions between algorithms and data explainable.

Explainability could help provide relevant information to domain experts necessary to recognise misconduct and hold responsible parties accountable (Diakopoulos, 2015). Explainability could help decision-makers justify their decision, which is often required for accountability. Having a human in the loop can enable decision-makers to explain and justify their decisions and use algorithmic outputs to address accountability concerns (Wieringa, 2020). Having to justify decisions could lead to a more motivated performance by the decision maker and a more interactive and effective exchange between humans and AI, also increasing their willingness to rely on the system (León et al., 2021). Moreover, knowing they will be held accountable

for their decisions, experts are more likely to spend additional time attending to AI and slowing their decision-making (León et al., 2021).

3.4 Building meaningful trust in AI

Trust in AI-driven decision-making can be defined as experts' willingness to rely on AI predictions (Yu et al., 2017). Algorithmic trust is usually built upon the understanding that the AI system and its outputs are fair, accountable, and transparent (Shin & Park, 2019). However, decision-makers relying on algorithmic predictions often do not have enough information about the system, which is needed to judge its trustworthiness (Ananny & Crawford, 2018). In many cases, their knowledge is limited to the official metrics, such as stated performance accuracy, and is insufficient to indicate trustworthiness (Yu et al., 2017). Poor knowledge about the system can lead to an overreliance on the prediction (automation bias) (Gaube et al., 2021) or unjustified rejection of it (algorithm aversion) (Dietvorst et al., 2015).

Meaningful trust is built upon an ability to understand and evaluate the accuracy of the prediction and make an informed decision on its reliability (Ribeiro et al., 2016). Thus, the goal should be to build meaningful trust by enabling a sufficient understanding of the system's workings. Explainability is often seen as a solution for promoting user trust in AI systems and their outputs. However, studies suggest that providing explanations can also lead to trust-related biases (Bussone et al., 2015). Explanations should not just create groundless trust and improve users' willingness to use AI but should enhance it (Liao et al., 2020). Ideally, the XAI techniques should build meaningful trust via explanations and appropriate interface design.

3.4.1 Automation bias

A lack of understanding of how AI systems work can lead to an overreliance on its predictions. This means that experts might be biased towards them and, as a result, fail to use their expertise to challenge AI outputs effectively. The HFE discipline has extensively studied automation bias phenomena. For example, Skitka et al. (1999) demonstrated how experienced pilots followed the autopilot system's advice incorrectly. They failed to react to irregularities and faults if the automated system did not detect them. Pilots also failed to properly assess the system's predictions, following them despite contradicting information from other sources (Skitka et al., 1999). In this case, automation appeared to reduce cognitive efforts put into

decision-making by experienced pilots. Unsupported trust in automation and its predictions has also been shown to arise when experts were not exposed to errors during the training process (Sauer et al., 2016).

Furthermore, the HFE literature shows that experts judge the trustworthiness of an automation system based on various heuristics when they cannot understand how it generates the output. This suggests that instead of using their expert knowledge or the knowledge of the systems, they rely on aspects that create a perception of the system being trustworthy. HFE research also demonstrates that people use the least cognitive effort required when relying on automation aids (Wickens & Dixon, 2007). When automation substantially aids their work, people use shortcuts rather than vigilant information-seeking and analysis strategies in their interactions with automation agents (Gigerenzer & Todd, 1999). As a result, they feel less of their action is required in the tasks, and automation becomes a powerful heuristic, leading to automation bias.

3.4.1.1 Automation bias towards highly reliable systems

Experts using automated systems also show bias towards automation if it performs consistently and reliably. They tend to become less vigilant and less likely to challenge reliable automation. In these cases, experts rely more on heuristic thinking than effortfully attending to and examining presented information (Parasuraman et al., 2000). For example, participants in a flight simulation study detected significantly more automation failures when automation reliability was low than when it was high (Bagheri & Jamieson, 2004). Studies exploring semi-automated driving found similar results (de Waard et al., 1999). Moreover, automation bias may be moderated by the historical reliability of (and associated trust in) AI tools in different fields. A study of 1,345 civil servants found no significant evidence of automation bias in their performance (Alon-Barkat & Busuioc, 2023). The authors argued that in areas where experts are used to reliable automation or AI (e.g., aviation, healthcare), they exhibit high trust in these systems and are, therefore, more likely to show automation bias than those in less AI-exposed domains, such as public administration (Alon-Barkat & Busuioc, 2023).

Extrapolations from HFE literature suggest that another heuristic that could lead to automation bias is the perceived superiority of AI systems. HFE research found that automated aids introduced as having superior analysis capabilities were less likely to

be challenged by experts (Lee & See, 2004). In such scenarios, users may interpret systems as more authoritative, making them more likely to over-rely on their recommendations. Similarly, in the study of civil servants, how AI was introduced to experts influenced their perception and willingness to rely on the system (Alon-Barkat & Busuioc, 2023). This perceived superiority of automation aid can result in pressure to conform to its "demands", even if they violate the user's sense of what is right (Parasuraman et al., 2000).

HFE research also shows that sharing decision-making tasks with automation can reduce effort as people may perceive themselves as less responsible for the outcome (Domeinski et al., 2007). In turn, a decreased sense of responsibility reduces efforts in monitoring and analysing other available information. Similarly, using AI-mediated communication between users was found to weaken the sense of responsibility of human communicators if things went awry, with responsibility instead shifting towards the AI system (Hohenstein & Jung, 2020).

3.4.1.2 Automation bias when the system's weaknesses are not known

Automation bias is also more likely if experts are not explicitly informed about a system's weaknesses. Studies in high-risk contexts such as aviation and intelligence work show that a lack of exposure to system errors during training or early system use could lead to overreliance in the future (Sauer et al., 2016). Aviation experts were less likely to complete necessary checks for verifying automated outputs if they were not trained using the system that exposed and explained errors to them (Mosier et al., 2007). This effect persisted even after they developed a good understanding of when verification was needed (Mosier et al., 2007). Similarly, Sauer et al. (2016) studied the effect of exposure to various automation failures on experts' trust. During the study, participants underwent training where they were exposed to either a completely reliable system or one of the three faulty systems: a) faults detected and reported, b) faults detected and not reported, c) faults not detected. Afterwards, they used a system that either failed to diagnose or misdiagnosed errors. Some participants showed signs of automation bias, i.e., they trusted system predictions more than their knowledge and, thus, failed to detect errors. This study revealed that users who overly trusted the system made more errors later if they were trained to use it with undetected faults but not if they were trained with the system that detected and reported mistakes. Extrapolating from these findings, it can be argued

that experts' trust in the AI system can lead to errors. However, making them aware of information underlying AI decisions (including its potential errors and limitations) through explainability while they are learning to use it could help promote experts' ability to evaluate the system outputs and build informed trust later.

3.4.1.3 Automation bias and expertise

The level of expertise can also be an essential factor leading to automation bias. Both low- and high-expertise users can over-rely on AI tools (Gaube et al., 2021). However, human-automation interaction studies suggest that novices are more likely to show automation bias than experts. A key factor may be how novices and experts approach tasks. For example, inexperienced weather forecasters apply a limited perspective on the task and use a fixed set of procedures to perform it (Stuart et al., 2007). In contrast, experts are more likely to be flexible in their decision-making and to interpret results and scenarios by relying on mental and conceptual models.

Moreover, highly experienced workers are more sceptical and seek confirmation of their expert opinion (Hilburn et al., 2014). Whether an output confirms or disconfirms an expert's opinion also plays a role. Experts are less likely to question outputs that align with their expert opinion—an example of confirmation bias—and therefore overrely on them in these contexts. For example, expert air traffic controllers were more likely to rely on automation if it matched their intuition or predicted outcome (Hilburn et al., 2014; Montgomery et al., 2004).

Users' self-confidence is another critical factor. More recent human-AI interaction studies showed that less experienced users have less confidence in their skills and are less flexible in their decision-making, so they are more likely to rely on automated systems as a default (Gaube et al., 2021; Nourani et al., 2020). Indeed, research suggests that user self-confidence may be a more critical determinant of reliance on AI outputs than confidence in the AI system itself (Chong et al., 2022). Finally, task familiarity is also essential. Users who are not experts in the domain but are highly familiar with the task also show overconfidence in their ability to perform it (Green & Chen, 2019a).

3.4.1.4 XAI potential in mitigating automation bias

XAI can improve understanding of how a system works and make users more aware of its accuracy, but providing explanations might not prevent automation bias (Bansal et al., 2021). Instead, applying XAI techniques and providing more information about

the AI's underlying logic have generally improved users' acceptance of the system and increased their trust in it (VanBerlo et al., 2021). Detailed explanations have also been shown to increase the risk of overreliance in the healthcare sector (Bussone et al., 2015). As Bussone and colleagues (2015) reported, very informative explanations simply created an impression that AI could determine even the most salient features that mattered for the diagnosis. In this case, users were more likely to assume that AI could use the same reasoning strategies as they would themselves (Bussone et al., 2015).

Moreover, not being suitable for the level of knowledge of interested stakeholders, XAI techniques might not be correctly used and interpreted, leading to overreliance (Kaur et al., 2020). For example, simply knowing the system is explainable has been shown to create an unjustified sense of trustworthiness and increase overreliance (Bansal et al., 2021). Explainability can also result in automation bias if explanations are used to identify the essential factors leading to AI predictions but do not explain why they were used. In this case, if the prediction matches the expert's intuition, they are more likely to accept it and demonstrate automation bias (Ghassemi et al., 2021). One way to overcome this overreliance was proposed by Buçinca and colleagues (2021b), who suggested using cognitive forcing intervention for people to engage with the AI-generated explanations more thoughtfully. The results of their study revealed that this approach can reduce overreliance compared to the standard XAI techniques. However, these techniques scored low on design acceptability and seemed more daunting to the study participants. This might suggest that there are instances in which explanations should be seamlessly embedded in decision-makers' workflow. For example, if the system has already been established as trustworthy through consistently accurate performance for a prolonged period. In that case, user-centred design might have to be prioritised. If explanations were always made to create friction in decision-making and encourage cognitive engagement, users might abandon the technology prematurely (Yang et al., 2019).

These findings suggest that explainability should be used cautiously to enhance trust rather than just improving users' willingness to use the AI system and as a panacea for algorithmic aversion. This is in line with Liao and colleagues' study (2020). Otherwise, using explainability to build trust in the system and its predictions might lead to unjustifiable confidence and overreliance (Kaur et al., 2020; Yeomans et al., 2019).

3.4.2 Algorithmic aversion

Experts also demonstrate distrust in algorithmic systems, systematically disregarding their predictions or refusing to rely on them (Veale et al., 2018). This phenomenon is referred to as an algorithmic aversion. It has been observed among lay users (Dietvorst et al., 2015) and experts, such as helicopter pilots (Veale et al., 2018). An ethnographic HFE study by Whalen (1995) of emergency dispatchers using a new automated dispatch system showed that they were reluctant to trust its outputs and continued checking them manually, even after six months following the introduction of the automation. More recent examples of algorithmic aversion by experts were observed by Lee and colleagues (2017), who studied automation practices in food donation services and interviewed various stakeholders participating in the process. One of the observations was that the community manager, making final food donation allocations based on the algorithmic analysis, continued using methods based on heuristics and logic that were adopted before the introduction of automation to make allocation decisions for 1.5 years. This might pose a risk of bringing implicit biases into the process if experts systematically reject AI predictions and override them when they do not align with their intuition.

3.4.2.1 Algorithmic aversion as a result of lack of explainability

Algorithmic aversion is often linked to a lack of explainability. For example, Brennen (2020) interviewed various stakeholders using AI-driven decision-making systems and found that most participants reported that they could not trust the AI system without understanding how it worked. Indeed, when AI-enabled tools are introduced, domain experts often display increased scepticism and distrust in algorithmic advice, even when these tools consistently outperform their judgment (Diab et al., 2011; Dietvorst et al., 2015). In some tasks, algorithms are accepted, while in others, they are prematurely rejected (Bogert et al., 2021). This decision can be determined by the specific features of the technology, such as the level of explainability (Barredo Arrieta et al., 2020) or display of accuracy metrics (Zhang et al., 2020). Various individual and contextual factors can also influence it. For example, more experienced users have been shown to rely on AI advice less than their novice colleagues (Logg et al., 2019; Povyakalo et al., 2013). Users who are less familiar with the task and those with less domain knowledge tend to rely more on AI recommendations, even when they are incorrect . However, experts have also been

shown to succumb to this bias and disregard predictions even when they are correct (Veale et al., 2018). In HFE research, experts have been shown to continue using old, ineffective decision-making methods even when highly accurate algorithmic systems are available (Lee et al., 2017; Whalen, 1995).

3.4.2.2 Algorithmic aversion as a result of a poor initial performance of AI

Algorithmic aversion could also result from a first impression of the AI system. For example, Hou and Yung (2021) conducted three decision-making experiments and found that users' reliance on AI systems depended on how their capabilities were framed compared to humans. The system with less power was more likely to be rejected. Dietvorst et al. (2015) showed that algorithmic aversion was particularly prevalent when system errors were displayed to the users before they made a reliance judgment. When participants were made aware of any errors made by the intelligent system, they were more likely to trust alternative predictions made by humans, even when these predictions were less accurate than the ones produced by algorithms. Yin et al. (2019) conducted a study where an AI system predicted speed dating outcomes, and participants were asked to compare and, if needed, adjust their predictions to the predictions made by the AI. They found that reported accuracy only had an effect if the observed accuracy of the AI performance was high. Otherwise, users tended to rely on their subjective observations more than provided metrics. Participants were more likely to rely on their predictions, even if they were less accurate than the ones produced by an AI system if the observed accuracy was much lower than the stated one. However, they showed that the initial effect of error disclosure can be reversed over time (Yin et al., 2019).

3.4.2.3 The agency can help overcome algorithmic aversion.

Algorithmic aversion can be overcome by allowing users to make minimal alterations to the AI. This way, a certain level of agency gives users a sense of control and increases trust in the system (Dietvorst et al., 2015). Algorithmic aversion can also be managed by introducing control tools that have little or no effect on the actual functioning of the system. For example, Vaccaro et al. (2018) showed that users felt more satisfied with their social media feed when they had an option to modify it, even when this option did not work. Sivaraman et al. (2023) conducted a study with clinicians using a conceptual prototype based on AI-generated treatment insights for type 2 diabetes medications. The authors concluded that allowing the medical

experts to customise the model made them more likely to rely on it. Factors such as the complexity of the task or organisational impacts are still underexplored when trying to understand algorithmic aversion (Mahmud et al., 2022). Dietvorst et al. (2016) conducted three experiments in which participants either relied on the AI suggestions or their own forecasts (e.g., predicting students' test scores), having varying levels of control in modifying systems parameters. The study revealed that users were more likely to rely on algorithmic forecasts if they could make changes. The authors concluded that users' trust in AI predictions could increase if they had some control over the system, even if it was restricted to making minimal alterations.

3.4.2.4 XAI as a solution for algorithmic aversion

Another way to reverse users' algorithmic aversion is to explain how the AI system works. Yeomans and colleagues (2019) examined how explanations affect users' trust. They showed that if participants received a description of how the recommender system worked, they felt more confident in it, and their aversion disappeared. However, understanding of the AI system was self-reported rather than measured in this study. The participants judged the system not on its accuracy but on how comfortable they felt using it. This suggests that trust in AI systems can be increased by the easy-to-comprehend explanations that give users a false sense of understanding and comfort. Thus, it is essential to challenge whether users understand the AI processes. However, as previously discussed, simply making users aware of any prior AI errors might diminish their trust in the system and make them more likely to trust less accurate predictions made by humans (Lee et al., 2017; Yin et al., 2019). Explanations must also be robust, as poorly structured and non-informative explanations can lead to experts' overreliance on themselves (Bussone et al., 2015). Explainability could increase the adoption of AI-driven systems in high-risk domains, such as the healthcare sector and pharmaceuticals, where any unexplained mistakes and unpredictable AI behaviours can mean the fast rejection of algorithmic solutions (Gilvary et al., 2019).

3.5 Challenges of staying in the loop

Research shows that experts often cannot stay in the loop when using AI-driven decision support systems. In AI-supported decision-making contexts, AI recommendations often do not improve the accuracy of experts' decisions (Jacobs et

al., 2021; Majid et al., 2011; Wang & Yin, 2021) or result in a worse performance than a human or an AI alone (Amann et al., 2022; Gaube et al., 2021; Micocci et al., 2021). These findings suggest that decision-makers using AI systems do not have meaningful control and might struggle to exercise their expert skills effectively. Thus, the challenges preventing domain experts' effective use of AI must be better understood in order to design usable and practical explainability solutions.

Research suggests that in many instances, decision-makers are restricted by the circumstances of the specific situation in which algorithms are embedded, such as time limitation, insufficient qualifications, or inadequate access to the relevant information necessary for meaningful human input to be possible (Ananny & Crawford, 2018; Shin & Park, 2019; Wagner, 2019). They might also lack a basic understanding of the system they use (Wagner, 2019; Young et al., 2019). HFE literature suggests that experts might feel cognitively overburdened by the new responsibilities of reviewing and evaluating automated system outputs (Sheridan, 2012). Extrapolations from HFE studies suggest that introducing AI systems might also disrupt experts' workflow and restructure their familiar sequence of tasks, complicating their use of expertise (Cork et al., 1998) and even leading to skill deterioration (Bainbridge, 1983).

This subsection looks at various aspects that challenge the expert ability to stay in the decision-making loop, such as i) the AI system design that does not align with experts' needs and workflows, ii) experts' lack of understanding of how the AI system works, iii) deterioration of their skills, and iv) increased cognitive load. This subsection is informed by the second stage of the literature review and HFE research literature exploring human-automation interaction dynamics.

3.5.1 Misalignments with experts' needs

One of the issues related to expert-AI interaction is that AI systems often do not support experts' needs. Interviews with 28 workers facing high job demands using productivity assistant AI tools showed that the view of users' work habits presented by these technologies did not match their experienced reality (Cranefield et al., 2023). The mismatch between the lived-in experience and technology representation required significant time and cognitive efforts for experts to understand the technology and improve its accuracy. Some users changed their work practices to incorporate AI into their workflows, while others resisted making changes and

instead rejected the technology. Similarly, the longitudinal study of an organisation in the international maritime trade found that introducing an AI system for data analysis and prediction work resulted in a redistribution of expert skills to augment and improve its accuracy (Grønsund & Aanestad, 2020). Another study conducted in an automated navigation context showed that while designers assumed that experts were in the loop, expert navigators referred to their changed role as a backup in case automation failed (Veitch & Alsos, 2022). Cases of poor technology fit in the workflow are also evident in the medical domain. Burgees et al. (2023) observed clinicians using new AI technologies and reported that they often failed to support clinicians' efforts to integrate information into practitioners' decision-making. For example, AI design did not consider different knowledge and resources (e.g., recent journal articles, other AI tools, and norms within a clinic/hospital systems) that clinicians were used to use. Another study reported that AI systems did not effectively support clinical processes, operational workflows, and practices. This led to inefficiency or low task performance in addressing patient needs, service quality, and satisfaction (Boch et al., 2022). In natural science research, AI adoption is often hampered by the gap between the knowledge experts acquired via academic training and the knowledge required in practice when AI tools are introduced (Ayres et al., 2021).

Human ability to stay in the loop can also be challenged if the AI system does not align with their level of expertise. This aspect has been shown to influence users' initial trust and acceptance of AI systems and their outputs (Nourani et al., 2020). For example, novice users often struggle to calibrate their trust based on the observed AI performance and overrely on algorithmic advice (Bussone et al., 2015; Dikmen & Burns, 2022; Micocci et al., 2021; Schaffer et al., 2019). Unjustified novice acceptance of technologies has been observed in the radiology sector (Gaube et al., 2021) and among immigration centre workers (Janssen et al., 2022). On the other hand, experienced decision-makers are often more sceptical about new technologies in their area of expertise (Nourani et al., 2020). Their perception of system accuracy is also more susceptible to first impressions. Observing errors early in the process can lead to the rejection of the AI system, whereas experiencing high system reliability can lead to future bias towards automation (Nourani et al., 2020).

Humans might be out of the loop if the AI system does not provide the information they need to make an informed decision. HFI research shows that automation can

deprive experts of critical feedback needed to assess the state of automation and its ability to perform tasks. For example, when automation was introduced to support operators in paper-making plants, their task changed from raw data processing to information integration (Lee & Seppelt, 2009). As a result, operators lost the information associated with informal feedback (e.g., smells and sounds), depriving them of contextual information that could help them diagnose automation failures and intervene effectively (Lee & Seppelt, 2009). Similarly, essential information from vibration and smell was lost in the automation of control operations in the aviation sector (Moray et al., 1986). Automating the auto-feathering systems in commercial aircrafts removed the signal informing pilots about engine shut-downs, leading to missed warning signs and incidents (Billings, 1991). The lack of system transparency or contextual feedback about its performance often becomes an issue during system failures as operators lack information for detecting or addressing them (Billings, 1991). When experienced weather forecasters had to adapt to automation, they felt that it made them less flexible and more passive, pigeonholing or disconnecting them from their preferred data analysis methods (Stuart et al., 2007). The feeling of confusion due to the loss of information has been shown to force experts to surrender to their old decision-making methods (even if less effective), for example, by manually searching for information (Lee et al., 2017). Similarly, AI recommendations can deprive users of important information, such as historical cases of medical diagnoses (Gu et al., 2020). Moreover, AI recommendations without additional information can be frustrating and demotivating to experts (Klein et al., 2006a, 2006b; Yang et al., 2019).

3.5.2 Disrupted experts' workflows

HFE research suggests that introducing AI systems might also disrupt experts' workflows and change how they approach their tasks, adding to experts' busy workloads and causing frustration (Elwyn et al., 2013). Experts often struggle to apply their skills when new factors, such as automation support, are introduced (Klein et al., 2006a). AI systems have also been shown to make experts feel a loss of control and agency (Yang et al., 2021). Failure to appreciate the context in which decisions are typically made without algorithmic support is one of the reasons why AI systems fail in expert contexts (Wagner, 2019). Poor contextual fit means decision-makers might feel limited and resist relying on a system's predictions (Khairat et al.,

2018; Yang et al., 2019). They also might lack the means or time to make an informed decision (Wagner, 2019). Interviews with public sector workers showed that how users interacted with AI systems and whether they relied on them were influenced by how well the system aligned with their natural workflow and organisational context (Veale et al., 2018). HFE research suggests that disrupting the decision-making workflow can prevent experts from using decision-making strategies learned from experience (Montgomery et al., 2004). Subsequently, experts, when introduced to automation systems, are likely to rely on their common sense or heuristics, usually searching for aspects confirming their intuition and failing to notice errors (Nickerson, 1998).

The introduction of AI systems can restructure experts' workflow, disrupt their familiar task sequence, and introduce new challenging tasks, such as reviewing AI outputs. This changes what strategies experts use, how they perceive information, and how they act in a specific context, potentially leading to ineffective use of their time and cognitive resources. HFE research suggests that experts do not just reduce the amount of work they do when certain parts of tasks are automated. Instead, they rely on different strategies to perform that task (Bainbridge, 1983). When automation introduces new tasks in operators' workflow, experts' familiar workflow structures can be disrupted (Cork et al., 1998). As a result, they might struggle to adapt their everyday work strategies and try to tailor the system or the task to accommodate the automation (Cork et al., 1998). When the system does not provide an option to be tailored, users might be forced to tailor their tasks, which could increase their workload (Cork et al., 1998). For example, physicians using automation aids learned to manipulate monitors displaying physiological data to fit their work strategies. However, because this manipulation was an additional task physicians had to perform, they avoided using the system in high-workload situations (Cork et al., 1998).

Workflow changes can also lead to difficulty following a task's familiar sequence of steps. Many tasks have sequential constraints, a set of actions that must be performed in order. Introducing AI systems can change the familiar sequence of tasks, for example, by removing a step of gathering data (Altmann et al., 2014). When the task sequence is disrupted, experts in health care domains have been shown to make errors and repeat the procedure, for example, by prescribing medication twice (Altmann et al., 2014). To perform a task correctly under sequential

constraints, the cognitive system has to keep track of where it is in the sequence and select the correct next step when one step is complete (Altmann et al., 2014). Changes in the structure of the task can make it difficult to follow the steps of the natural sequence. For example, automation research showed that operators' reactions are slower and less integrated when they cannot generate the activity sequence themselves (Janssen et al., 2015). Not having a task structure to follow also disrupts experts' ability to monitor their progress. Under manual control, they obtain information about the results of their actions and then can correct themselves (Smith, 1979). Without this information, they are more likely to repeat the same type of errors (Wiener & Curry, 1980). Interruption of the task can also interrupt the user's thought processes (Altmann et al., 2014) and initiate a switch between tasks requiring time and cognitive resources, negatively affecting their performance (Janssen et al., 2011). Long and complex interruptions can be particularly costly (Mark et al., 2008, 2012; Monk et al., 2008). Moreover, interruptions can also break the expert's flow state (Taekman & Shelley, 2010).

3.5.3 Increased cognitive workload

When AI is introduced in the decision-making workflow, an expert role often shifts from production to evaluation. When automation was introduced, manual control tasks turned to monitoring tasks, leaving humans to supervise the automation (Sheridan, 2012). However, monitoring can impose a considerable workload (Grubb et al., 1995; Warm et al., 2008). For example, in the aviation context (e.g., detection of air traffic in an aircraft's vicinity), pilots' workload was not reduced but was shifted to performing different tasks such as supervising. Pilots were spending more time interacting with automation and trying to understand it instead of concentrating their efforts on their primary task of flying the aircraft (Rudisill, 1995). In other domains, operators supervising automation also spent significant time and effort learning how to manage the new technology (Baxter et al., 2012). In the context of automation, the workload is further increased because users' situational awareness is reduced, as well as their perception of data and elements of the situation, comprehension of the situation, and the projection of future status (Endsley, 1995; Manzey et al., 2012; Metzger & Parasuraman, 2001). Low situational awareness can significantly decrease experts' ability to effectively monitor and observe automation errors and to

determine whether the given situation was outside the bounds of automation capabilities (Jones & Endsley, 1996).

The mental load might also increase as the complexity of automation and AI increases. It can be challenging to understand and anticipate complex system's behaviour. For example, when traders in the digital stock exchange changed roles from executing trades to monitoring the automated process of trading, they underperformed as they could not effectively monitor the trades in real time (Haldane & May, 2011). As a result, they resorted to watching them at a higher level of abstraction. They required additional resources to process that information, thereby missing more trades executed (Haldane & May, 2011). Supervising automation, including AI systems, is further challenged by the opacity of these systems. More features and modes create more possible interactions among system components while reducing the system's predictability as the system increasingly considers multiple factors or component states (Endsley, 2023). This can lead to unfamiliar and infrequent system states, which add to the challenge of comprehending systems' workings. For example, even well-trained pilots were overwhelmed by the unexpected behaviours of complex flight automation systems (Wiener & Curry, 1980).

Having to evaluate automation outputs during cognitively demanding tasks has been shown to increase mental workload (Lee & Seppelt, 2009). This phenomenon has been termed "clumsy automation" in HFE research literature (Cook & Woods, 1997). For example, automation has been shown to reduce pilots' mental workload when it is already low during easy tasks, as when the plane is on autopilot during a straight flight. However, automation increased the mental workload of pilots when the flight-related workload was already high, e.g., during landing, as they had to simultaneously reprogram the system managing autopilot, activate landing procedures, and manage communication (Wiener & Curry, 1980). Humans are also ineffective in shifting cognitive resources saved by automation to support more complex tasks. In Metzger & Parasuraman (2005), air traffic controllers used automation designed to aid conflict detection and resolution tasks. This was expected to free up enough mental resources that controllers could allocate to performing more complex tasks. However, automation did not reduce the cognitive workload in demanding tasks, such as communication and accepting and handing off aircraft. Either the aid did not free enough resources, or the controllers could not

allocate them to improve communication performance. Studies exploring AI-driven decision-making to support government tasks showed that the new technology often only reduced the easy assignments but left the difficult ones to the government workers, making their work more complex and fragmented (Lindgren, 2023). People experiencing high workloads have been shown to regress from concurrent performance (i.e., managing several tasks at once) to a sequential mode of multitasking (Wickens et al., 2013). Sequential mode requires a person to switch attention from one task to another and make implicit decisions to perform some tasks while sacrificing or ignoring others (Raby & Wickens, 1994). Switching between tasks requires additional cognitive effort, which creates this vicious cycle of increased task demands. The HFE studies in automation-aided aircraft control showed that multitasking under a high workload led to complacency and failure to monitor automation and detect faults (Parasuraman et al., 1993). When flight operators' attention was distributed over multiple sources, they could not effectively attend to all of them (Parasuraman et al., 1993). Operators experiencing high task demands have also been shown to sacrifice difficult tasks over easier ones (Gutzwiller et al., 2014) or use shortcuts in performing one of the tasks (Wickens et al., 2013).

3.5.4 System opaqueness

Experts working with AI systems are often not informed on how they work and are unaware of their capabilities and weaknesses (Young et al., 2019). Young et al. (2019) observed public sector workers using AI-driven systems for surveillance and reported that they lacked knowledge about the systems they used and drastically underestimated their complexity (Young et al., 2019). Similarly, Veale et al. (2018) interviewed public sector workers using predictive systems and discovered a potential disconnect between organisational and institutional outlooks and realities interfering with their ability to use them effectively. The authors suggested that people directly involved in using these systems should be included and better informed when attempting to apply algorithmic methods more fairly and transparently. Without a deeper understanding of a system's inner workings, decision-makers find it difficult to determine whether they should rely on algorithmic outputs (Yu et al., 2017). Wagner (2019) analysed human agency in decision-making in three cases: self-driving cars, border searches based on passenger name records,

and content moderation on social media. The author showed that the methods and regulations used to include human agents in decision-making were ineffective. A decision-maker is considered out-of-the-loop if unable to identify irregularities and errors in the system and take corrective action (Rahwan, 2018).

When experts cannot understand AI outputs, they are more likely to rely on them when they are incorrect, performing worse than without any recommendations from the AI system. For example, when an AI system was used to support emergency call dispatchers, AI on its own showed a better ability to detect cardiac arrest from the phone call recording than a human expert on its own or in collaboration with a system. Dispatchers could not make informed decisions about whether or not to trust the system and comply with its recommendations (Amann et al., 2022). Micocci et al. (2021) studied clinicians making referrals using a fictitious AI-based decision support tool. They found that physicians could not identify and overturn AI recommendations when they were erroneous and still relied on AI. They were especially likely to rely on AI outputs that aligned with their opinion, succumbing to a confirmation bias. Those with less experience in performing the specific task and lower domain knowledge are also likelier to overestimate AI's accuracy (Gaube et al., 2021). Gaube et al. (2021) provided radiologists with chest X-rays and either accurate or inaccurate diagnostic advice from a human expert or AI (humans generated both diagnoses). Those with more task expertise were likelier to judge AI-informed advice as lower quality than those with less experience. However, all participants' diagnostic accuracy was significantly worse if they received incorrect advice independent of the source. Furthermore, AI recommendations have been shown to make people doubt their expertise, nudging them to accept AI recommendations over their expertise, even when the latter is correct (Burkart & Huber, 2021).

System opaqueness also reduces situational awareness and affects experts' ability to evaluate system outputs. HFE research suggests that the system's complexity and opaqueness can make it more difficult for users to create an accurate mental model of the system needed to interpret the information provided correctly (Baxter et al., 2012). A recent study conducted in local government organisations with workers using automation aids showed that monitoring tasks were barely possible from the human cognition point of view (Lindgren, 2023). AI systems functioned faster than experts could follow, and the actual workings of the system were not transparent to them (Lindgren, 2023). Monitoring is particularly demanding

when automation functions unreliably. For example, Metzger et al. (2005) studied the performance and workload of air traffic controllers using reliable and inaccurate automation when making aircraft-to-aircraft conflict decisions. When it functioned unreliably, traffic controllers were better at detecting conflicts without it. They could not monitor the automation effectively or transfer their skills if they had to observe decisions made by systems they could not understand.

3.6 XAI challenges in expert contexts

Often, an AI output without a comprehensive explanation does not give the expert enough information to make justifiable decisions (Amann et al., 2022). Experts cannot effectively integrate the AI outputs into their decision-making without understanding them. However, despite continuous research efforts and advances in XAI research, many proposed technical explainability approaches lack usability when implemented in practice (Abdul et al., 2018; Leichtmann et al., 2023). For example, Bhatt et al. (2020) interviewed data scientists and other stakeholders across 30 organisations. The study revealed that explainability was mainly viewed as a tool for debugging the model used by software engineers. Interviewed decision-makers did not see explainability as a valuable feature and felt that it did not represent or meet their needs. It has been shown that explainability often fails to improve decision-making. In some instances, it can have undesirable effects and mislead experts (Jacobs et al., 2021).

Moreover, explainability approaches that are too complex or not tailored to expert information needs can also fail to motivate them to attend to the provided explanations (Eiband et al., 2018). Without adding a clear value to the expert work, explainability could become a formality and be considered a redundant feature (Schemmer et al., 2021). This section explores the critical challenges with current XAI approaches in supporting the human-in-the-loop approach.

3.6.1 Explanations do not align with expert reasoning

Explainability solutions are often designed with the assumption that experts have a certain level of data science or computational knowledge and skills (Tomsett et al., 2018). Only in recent years have researchers started paying attention to decision-makers' needs to ensure that explainability is usable to them when applied in practice (Millecamp et al., 2019; Rosenfeld & Richardson, 2019). Thus, XAI

techniques are often too complex and incomprehensible to domain experts (Conejero et al., 2021; Woodruff et al., 2020) and ineffective when introduced in actual work scenarios (Anjomshoae et al., 2019; Hoffman et al., 2018). When explanations are presented as complex visualisations or statistics, specific skills are required to interpret them correctly (Hadash et al., 2022). However, domain experts are often not required to have these skills as part of their job. For example, Conjero et al. (2021) explored the effectiveness of data visualisation in various governmental decision-making situations. The authors used interactive dashboards, charts, maps, and diagrams to illustrate patterns, relationships, and correlations in data. These visualisations exposed data points a human would not otherwise pick up. However, the visual explainability design was not accessible to the public administrators in the education and employment sectors (Conejero et al., 2021). If attending to explanations requires learning non-domain-specific information, decision-makers are unlikely to use them. For example, physicians have been shown to be unable or unwilling to remember information unrelated to their domain due to their already intensive workload (Gu et al., 2020). The requirement to interpret complex explanations can also increase the potential for an error and add to the responsibilities of experts (Bertrand et al., 2022).

Explainability might also be ineffective if it does not represent experts' decision-making habits and familiar problem-solving strategies. When AI was introduced to support decision-making in diagnostic medicine (e.g., predicting breast cancer), physicians found it challenging to integrate its predictive outputs without any supporting contextual information (Gu et al., 2020). They reported that AI systems disrupted their decision-making ability using historical medical cases, as they could not access this information through AI outputs. Explanations that did not fulfil this requirement were unhelpful and only added to their workload (Gu et al., 2020). Decision-makers have been shown to ignore explanations that are difficult to contextualise and integrate with their domain knowledge (Bansal et al., 2021; Naiseh et al., 2021). If explanations are unhelpful and perceived as time-consuming, experts are likelier to reject or inspect them superficially (Bansal et al., 2021; Naiseh et al., 2021; van Berkel et al., 2019). Explanations are also likely to be ignored if they are too simplistic (e.g. if they repeat experts' existing knowledge) (Bussone et al., 2015; Naiseh et al., 2021). When healthcare experts received explanations that lacked medical detail, they did not use them (Bussone et al., 2015). Explanations that are

simply available but not helpful in promoting understanding can also feel like a time waste and cause frustration (van Berkel et al., 2019).

3.6.2 Explanations can lead to biased decision-making

Explainability not designed for experts can be overly complex and difficult to interpret. Even when incomprehensible, it can still be perceived as a sign of trustworthiness. Interviews with data scientists using popular explainability techniques revealed that these techniques were often misused by users and led to their overreliance on AI predictions (Kaur et al., 2020). Explainability can enable specific heuristics about the AI system. For example, Bansal et al. (2021) asked participants to solve a task with the support of AI with accuracy comparable to humans, and some conditions also provided explanations. The authors observed that explanations did not improve the accuracy of human-AI teamwork. Instead, explanations increased users' willingness to accept AI's recommendations regardless of their correctness (Bansal et al., 2021). These findings suggest that users might perceive explainable AI as more trustworthy because it explains its actions.

When XAI is seamlessly embedded in the system design, it can lead to overreliance (Buçinca et al., 2021b). Buçinca et al. (2021b) provided participants with AI-generated explanations, but in some conditions, they used interventions (design friction) encouraging them to engage with the provided explanations cognitively. Participants who received explanations without interventions were more likely to overrely on the incorrect recommendations than those who received no or cognitively engaging explanations. These results suggest that it might be less risky not to use explanations at all than to use explanations that are not engaging. When users do not attend to explanations or do not understand them, they interpret their validity based on heuristics, such as how much information was provided. Bussone et al. (2015) found that detailed and informative explanations that included all items from medical history, symptoms, and examination results increased trust in the system and its outputs, increasing overreliance. Informative and detailed explanations led healthcare experts to believe that the system used the best available medical knowledge and reasoning processes similar to theirs. Explanations can make experts more compliant with the algorithmic systems (Naiseh et al., 2021). Explanations can also set inaccurate expectations for the system. They can make it

feel superior to the user, also increasing the potential for overreliance (Kocielnik et al., 2019). An unjustified sense of confidence in the system can make the user feel that the AI system is fairer than it is, making them less likely to challenge it (Green & Chen, 2020).

Explanations that do not fit with the expert's role, their cognitive state, and the contextual constraints 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 (van der Waa et al., 2021). Van der Waa et al. (2021) used explanations in a medical triage task. They asked healthcare professionals to make decisions under time pressure and with insufficient resources available to comply with all the medical guidelines. They found that experts primarily did not attend to explanations during the task, either because of the time pressure or the unjustified trust in the system. This study showed that meaningful control by experts is questionable even when decisions are sensitive and when experts receive explanations. In a different study, van der Waa et al. (2021) provided participants with rule- and example-based contrastive explanations for the AI-supported diabetes self-management (i.e., personalised AI advice on aspects such as a lower insulin dose based on the current temperature). They found that both types of explanations persuaded users to follow the incorrect AI advice without improving the overall performance. Authors argued that users did not have the means to effectively use explanations, as they did not provide any information about the underlying rationale or causality (van der Waa et al., 2021).

3.7 Chapter overview

The expert role in AI-supported decision-making is essential, especially when AI is used to make sensitive decisions. However, experts often lack the means to form a meaningful trust in AI outputs and to make informed decisions based on them (Raghu et al., 2019). In many cases, humans are not in the decision-making loop, and when they are, they follow various heuristics to judge the AI outputs. Suppose humans cannot effectively oversee the AI outputs, the risk of biased decisioning increases, as well as the risk of unfairness towards the affected individuals, such as patients waiting for a diagnosis. Attempts to make algorithms transparent so experts could inspect the available information are either overwhelming or too simplistic and fail to provide relevant information that could help them inform their decisions

(Bussone et al., 2015). If experts do not have sufficient means to evaluate AI outputs, their accountability for the outcome is also questionable (Diakopoulos, 2019).

Experts' ability to stay in the loop is challenged by the lack of information that they could use to judge the trustworthiness of the AI outputs. This leads to heuristic thinking and trust-related biases in decision-making. Experts are likely to overrely on AI outputs when they are unaware of the strengths and weaknesses of the system or have not been exposed to any potential errors while trying the system (Sauer et al., 2016) or when it aligns with their judgement (Hilburn et al., 2014).

On the other hand, if experts cannot effectively evaluate AI outputs, they might reject them without proper consideration (Dietvorst et al., 2015) or continue relying on their old decision-making methods, missing out on the potential benefits of using AI (Lee et al., 2017). The rejection of AI recommendations is often linked to a lack of explainability and is particularly likely if the system underperforms during the initial period of trialling the technology.

Extrapolations from the HFE research suggest that experts' ability to stay in the loop can also be challenged if the AI system design does not align with their needs and deprives them of relevant contextual information or performance feedback (Lee & Seppelt, 2009). Experts' ability to effectively use AI systems could be further complicated because of the increased cognitive workload due to the need to supervise AI performance, review its outputs, and adapt their workflows to accommodate the system’s capabilities and limitations (Bainbridge, 1983; Sheridan, 2012).

However, more than simply providing explanations is needed to help experts overcome these challenges. Explanations are ineffective if they do not align with expert reasoning (Conejero et al., 2021) or require additional skills to interpret them (Bertrand et al., 2022). Ineffective explanations can distract users from fairness issues (Binns, 2018) and lead to overreliance on the system and its outputs (Bansal et al., 2021; Buçinca et al., 2021b). On the other hand, explanations not tailored to experts' cognitive and contextual needs are likely to be ignored because they are either seen as unnecessary or too demanding (Bussone et al., 2015; Gu et al., 2020).

This chapter indicates the need for tailored explainability that would effectively inform experts and align with their decision-making strategies, workflows, and

reasoning patterns. It also shows that explainability should feel valuable to experts, for example, by applying design interventions that would motivate them to engage with explanations. Chapter 5 will address the need for explainability to match expert cognitive and contextual needs. Chapter 6 will address the need for explainability to be engaging, motivating and stimulating.

Chapter 4 Foundations for effective explainability: Strategies to increase the acceptability of AI through collaboration and ongoing support

Chapter 4 discusses the barriers to effective AI adoption that need to be addressed before explainability is introduced to experts. In this chapter, I propose four steps that should be taken to help experts successfully adopt and use AI systems: i) develop collaborative training solutions, ii) actively support experts during the initial stages of system use, iii) clearly communicate the capabilities of a system, and iv) follow predefined collaboration rules. These steps are intended to promote collaboration between AI developers and domain experts and lay the foundation for explainability as a useful feature in experts' workflows. This chapter is informed by a contextual enquiry study and semi-structured interviews with science experts and AI developers.

This chapter explores *explainability foundations*, first of the three properties of the holistic XAI approach proposed in this thesis.

4.1 Chapter introduction

When researching explainability and AI in expert domains, it became evident that, in many cases, experts are unable to include AI in their workflows or do not adopt these systems effectively. In less technical domains, there are very few research cases where explainability is a feature actively being used or sought after by experts. One of the potential reasons is that experts often fail to reach the point where they can use AI efficiently enough for explainability to be needed and helpful. Research analysing explainability and human-AI interaction often explores users' actions and feelings towards AI outputs and explanations or studies how certain software features affect users' trust and reliance on AI. However, the more practical hurdles for low AI adoption are poorly understood. There is a need to recognise the barriers to AI adoption in order to lay the foundations for explainability.

4.1.1 Expert-AI interactions in life science research

In this chapter, I explore AI adoption in a life science context where expert scientists have increasingly depended on AI tools for data analysis. AI is intended to automate mundane, repetitive tasks so human expertise could be directed to higher-level and

creative activities (Janssen et al., 2019a; Sousa et al., 2019; Zhang et al., 2020). In many cases, AI is also essential to coping with the vast amount of data available due to technological advances (Haenssle et al., 2018). Complex AI systems and the availability of biological data have already contributed to significant achievements in fields such as Alzheimer's research (Zhao et al., 2024), drug discovery (Paul et al., 2021), genomics (Caudai et al., 2021), and diagnostics (Kumar et al., 2023). However, without AI being appropriately implemented, there is a risk of "bottleneck" situations in domains within the natural sciences (Shani et al., 2023). If scientists are constrained to manual data processing, valuable data is left unused, and significant scientific discoveries are slowed down or completely halted (Marx, 2013). Domain experts acknowledge the benefits AI systems offer (Woodruff et al., 2020). Still, in practice, they remain reluctant to integrate AI systems into their workflows (Dietvorst et al., 2015), they use them ineffectively (Jacobs et al., 2021; Majid et al., 2011; X. Wang & Yin, 2021), or their performance worsens with the support of AI (Diab et al., 2011; Dietvorst et al., 2015; Green & Chen, 2019a; Yin et al., 2019).

4.1.2 HCI response to low AI adoption

The HCI community have approached this issue of AI resistance by proposing more usable, accurate, and explainable technologies (Holzinger et al., 2017; Ribeiro et al., 2020; Samek & Müller, 2019). Increasing efforts have been directed at understanding how model transparency influences the adoption of AI systems in practical settings, for example, by studying physicians' interactions with AI (Dey et al., 2022). These efforts are undoubtedly needed to develop ethical and user-centred technologies. However, domain experts prematurely reject and misemploy AI systems, even if they are explainable and outperform human experts (Diab et al., 2011; Dietvorst et al., 2015). Despite the growing user-centred research focus, experts' needs continue to be misinterpreted in these systems' design, leaving them frustrated with technologies that do not fit their preferences, knowledge, and workflows (De-Arteaga et al., 2020). I argue that more fundamental barriers exist between experts and AI tools designed to support them, which are currently underexplored. First, there is a lack of research examining how and why experts' needs are misinterpreted by those developing and introducing AI-driven technologies. Second, outside of studies surveying organisations and looking into higher-level barriers to AI adoption (e.g., environmental, organisational) (Alsheibani

et al., 2019; Pan et al., 2022), few have explored the emotional, contextual, and collaborative obstacles that might have to be removed before proposing more usable, explainable technologies to domain experts.

4.1.3 Study aims

Rather than exploring the technology features influencing AI acceptance, I focus on understanding i) the practical and contextual barriers that experts face when AI systems are introduced into their work practices and the processes that happened before the introduction of AI that could influence these barriers. I also explore the AI practitioners' role and observe ii) their perceptions of the reasons for low AI adoption among experts and iii) their approach to this issue. I then iv) compare practitioners' and experts' perceptions and follow how inaccurate assumptions can stifle effective AI adoption. Lastly, I study v) how effective collaboration between domain experts and practitioners is reflected in experts' motivation to use AI systems. This chapter reports on an ethnographically informed study of 10 in-depth interviews with AI developers and natural scientists working at the same organisation, which faces low adoption of algorithmic systems and bottleneck issues, with more data being collected than possible to process manually.

4.1.4 Study contribution

The analysis revealed how miscommunication and frustration between practitioners and experts can stifle AI adoption. In this chapter, I show that practitioners overestimate the technical knowledge of expert scientists and create an environment in which experts are expected to manage aspects such as software installation and command lines independently. The study results suggest that the lack of support in performing these simple tasks can put an emotional burden on experts, take their attention away from their projects, and make them averse to AI systems. Furthermore, I show that experts are likely to reject an unfamiliar system if its capabilities are not communicated in a way that is relevant to their project. In contrast, they are likely to invest time in exploring them if they understand how system capabilities align with their needs. However, practitioners focus on communicating the technical parameters of their systems, expecting experts to discover how they can use AI to their benefit independently. Finally, this chapter reports positive cases showing that continuous and goal-driven collaboration can

help resolve these miscommunications and develop mutual understanding between teams representing different disciplines.

The chapter makes three contributions:

1. It provides AI acceptance requirements that could help to overcome the obstacles to effective AI adoption. These are: i) developing collaborative training solutions, ii) actively supporting experts during the initial stages of system use, iii) clearly communicating the capabilities of a system, and iv) following predefined collaboration rules.
2. It presents recommendations for effective multi-team collaboration that would support the implementation of AI systems in workplaces. These findings could inform practices in various domains and be generalised beyond the case organisation. The listed steps and collaboration rules could help organisations enable effective AI adoption, scaffold the development of usable technologies, and allow experts to benefit from the AI systems developed to support them so that they could focus on their expertise.
3. It describes the emotional and practical consequences arising from the sudden increased pressure for domain experts to adopt AI tools in their work and the lack of transitional support they receive. It also highlights the misappreciation of experts' relatively limited computational skills. These findings could inspire a shift in the research of AI applications in expert contexts.

4.1.5 Publications and impact

The study results reported in this Chapter were used to inform the report *Explainability in Expert Contexts: Challenges and Limitations in Supporting Domain Experts in AI-driven Decision-making*. I researched and wrote the report as part of the Department for Culture, Media and Sport Junior Policy Fellowship, and it is being published. The results and the proposed solutions will be shared with the organisation where the study was conducted to influence their practices directly.

4.2 AI in experts' workflows

AI should support experts by automating time-consuming and mundane tasks and allowing them to focus on exercising their expert skills (Zanzotto, 2019), but it often has the opposite effect (Green & Chen, 2019b; Raghu et al., 2019). Human-automation interaction studies call this phenomenon the "irony of automation" (Bainbridge, 1983; Janssen et al., 2019b). Researchers studying automation have shown that introducing automation can disrupt experts' workflows and change how they approach their tasks (Klein et al., 2007), adding to their busy workloads and causing frustration (Elwyn et al., 2013). Experts also report feeling a loss of control and agency when relying on automation or AI (Klein et al., 2006a; Veitch & Andreas Alsos, 2022; Yang et al., 2021). For example, experienced weather forecasters reported that automation made them less adaptive and passive, pigeonholing and disconnecting them from their preferred way of data analysis (Stuart et al., 2007). Feeling disengaged can force experts to return to more familiar but less effective methods, such as manual information search (Lee et al., 2017).

Experts often need to change their work methods to accommodate technologies that do not fit their workflows. Interviews with 28 workers facing high job demands using productivity assistant technologies showed that the view of users' work habits presented by the technology did not match their experienced reality. This mismatch required significant time and cognitive efforts for users to adapt (Cranefield et al., 2023). A longitudinal study of a maritime trade organisation that introduced algorithmic support for data analysis and prediction work found that technology resulted in a redistribution of expert skills to augment and improve the algorithm's accuracy (Grønsund & Aanestad, 2020). AI's disruption of experts' workflow is also evident in the medical domain. Burgess et al. (2023) observed clinicians using AI systems and reported that they did not support integrating information into practitioners' decision-making. Designers and developers failed to acknowledge different knowledge and resources (e.g., journal articles and norms within a hospital system) that clinicians use. Another study reported that AI did not effectively support clinical processes and practices, leading to inefficiency or low task performance in addressing patient needs, service quality, and satisfaction (Boch et al., 2022).

4.2.1 Algorithmic aversion

Experts often display scepticism and distrust toward algorithmic advice, even when AI systems consistently outperform human judgement (i.e., they demonstrate algorithmic aversion) (Bogert et al., 2021; Diab et al., 2011; Dietvorst et al., 2015). This effect can be influenced by the specific features of the technology, such as explainability (Barredo Arrieta et al., 2020), accuracy metrics (Zhang et al., 2020) and various individual and contextual factors. For example, more experienced users have been shown to rely on AI advice less than their novice colleagues (Logg et al., 2019; Povyakalo et al., 2013). Hou and Jung (2021) conducted three decision-making experiments and found that users' reliance on algorithms also depended on how powerful the algorithm was compared to humans. The system with less power was more likely to be rejected. Several studies suggest that having some control over the AI outcome could mitigate users' algorithmic aversion (Dietvorst et al., 2015; Sivaraman et al., 2023; Vaccaro et al., 2018). For example, Sivaraman et al. (2023) conducted a study with clinicians using a conceptual AI prototype for type 2 diabetes medications and concluded that allowing experts to customise the model increased reliance.

4.2.2 AI adoption

Barriers to AI adoption are less researched than algorithmic aversion. Automation studies suggest that experts were unlikely to rely on automation that lacked transparency, especially if its suggestions misaligned with their judgement (Westin et al., 2016) or if it jeopardised their agency (Bekier et al., 2012; Kaber & Endsley, 2004; Parasuraman & Riley, 1997). Langford et al. (2022) studied how increasing automation affected manufacturing controllers' views. Authors reported that experts viewed the lack of system understanding/insufficient training, skill degradation and changing roles as their biggest worries. The low acceptance of automation has been shown to result in the disuse of a system, failure to benefit from it, and diminished overall performance (Parasuraman & Riley, 1997). Studies of AI in high-risk contexts show similar results. In a clinical setting, medics were unwilling to integrate AI tools into their workflows if they perceived them as threatening their clinical autonomy or if the clinical value of these tools was unclear (Henry et al., 2022). Dorton et al. (2022) studied intelligence workers using AI and found that they lost trust, and either

reduced frequency or altogether ceased using AI after it underperformed. Several studies explored the issues of AI adoption at the organisational level. In a clinical setting, medics were unwilling to integrate AI tools into their workflows if they perceived AI as threatening their clinical autonomy or if the clinical value of these tools was unclear (Henry et al., 2022). Low AI acceptance by clinicians is considered a critical barrier to effective AI use for improving patient outcomes (Sivaraman et al., 2023). Alsheibani et al. (2019) conducted a questionnaire study involving 207 different-sized organisations about the main barriers to AI adoption. They reported environmental barriers, such as regulations, organisational obstacles related to a lack of top management support, and technical barriers, such as a lack of AI skills and employees' fear of change. Pan et al. (2022) studied the challenges of AI adoption in recruitment. The authors emphasised reducing obstacles to AI adoption rather than just developing practical AI techniques for the future of AI technology. They also drew attention to AI practitioners' efforts to understand the difficulty of using AI technologies with little or no technical background. They suggested that reducing complexity and making AI tools user-friendly should be prioritised to increase AI adoption. Factors such as the complexity of the task or organisational impacts are also underexplored when trying to understand AI adoption (Mahmud et al., 2022).

4.2.3 Theories of technology acceptance

The two most widely accepted theories of the user adoption of technology are the Technology Acceptance Model (TAM) Venkatesh et al., 2012) and the Unified Theory of Acceptance and Use of Technology (UTAUT) model (Kulviwat et al., 2007, Venkatesh et al., 2012). TAM focuses on the relationship between perceived usefulness and ease of use of new technology, predispositions towards the latest technology and behavioural intentions to use the technology. An individual's decision to engage or not engage in the behaviour of using a new technology is thus based on the expected outcome of using the technology and whether using a new technology is free of effort. The UTAUT model extends the TAM approach by proposing that technology acceptance depends on how widely others accept it by observing others within one's social interactions and experiences (Schunk, 2012; Venkatesh et al., 2012).

I argue that current research efforts exploring AI-expert interactions overlook the aspects of perceived usefulness and required effort. There has been little effort to examine practical hurdles and workflow changes that might prevent experts from using newly introduced AI systems. There is insufficient understanding of the barriers users face before and during the interaction with AI and how they perceive potential future benefits that could outweigh these barriers. In line with Pan and colleagues (2022), I argue that the importance of burdens related to domain experts having little or no technical background is underexplored. Research shows that experts understand that they need AI to advance their work (Shani et al., 2023). However, there seems to be a hurdle blocking users from adopting and effectively using AI systems. It is essential to understand the factors that impede users before trying to enhance the appeal of the technology.

4.3 Methodology

4.3.1 Contextual inquiry

The contextual enquiry method was chosen for the data collection (Raven & Flanders, 1996). This ethnographically informed method involves the observation of users' daily activities in their natural environment, such as their workplace. During the observations, users are also asked to explain why and how they perform certain actions. This method was chosen to help understand the practical expert-AI interaction and adoption challenges that laboratory or online studies might not capture. The contextual inquiry method can help gain a robust understanding of work practices and behaviours and recognise low-level details that have become habitual and ‘invisible’ to the user. It can also help reveal interruptions and illogical processes that influence human-AI interaction (Raven & Flanders, 1996). The purpose of contextual inquiry is to go beyond the interface design and understand users’ work processes. In cases where users were not supposed to be distracted or interrupted, a direct observation method was used. I would silently observe user behaviours, take notes and ask questions after the user completed their work task. The contextual inquiry study was conducted with five expert scientists using in-house AI software.

4.3.2 Semi-structured interviews

I also conducted in-depth semi-structured interviews with four practitioners (in-house software developers), the AI team lead and science director and five expert scientists

(the same experts that were observed using the contextual inquiry method) (Table 1). During the interviews, some practitioners showed relevant material (guides) and software to illustrate their answers, but this was not a planned part of the interviews, and, therefore, the notes from observing this were used only as additional enrichment during the data analysis. The interviews were guided by the Grounded Theory Approach, using the Theoretical Sampling Method (Birks & Mills, 2022). First, the initial interview with the AI team lead and science director revealed potential communication issues during the software introduction and showed that practitioners had assumptions about the reasons for low software adoption and expert expectations for it. This led to the inclusion of the additional questions in subsequent interviews. The interviews with experts were conducted after their attempts to use the in-house software, and additional questions were included depending on the observed behaviours.

4.3.3 The case study

The research was conducted in a large natural science research centre where the AI-enabled technologies are developed by the in-house software engineering team for the scientists within and outside the organisation. The organisation was recognised as a potential case study during the conference event, where they presented their work and acknowledged the issues of AI adoption by their scientists. After being approached by the principal researcher's supervisor, an informal meeting with the AI team lead and science director was organised, and the invitation for the principal researcher to visit the organisation to conduct the study was made. The organisation was chosen for the study because i) AI systems were essential for their experts' progress (as the volumes of data are too large to process manually), but the adoption of the AI-enabled software remained low, ii) the advances in data collection technologies were recent, and experts were new in having to use AI systems, iii) practitioners and experts worked in the same building, and some projects required their collaboration.

4.3.4 The research aims and questions

1. **Investigate the practical and contextual barriers that experts face when new AI systems are introduced into their workflows.**

a. What practical difficulties do experts experience when new systems are introduced to support their work?
b. How do experts respond to these practical difficulties? What strategies do they use to overcome them?
c. What contextual factors influence experts' ability to use new AI tools?
d. What strategies do experts use to respond to different contextual factors that interfere with or support their use of new AI tools?

2. **Identify the challenges to successful AI use and adoption occurring in different stages of AI system introduction, including before experts start using them.**

a. How does the way the new AI system is introduced to experts affect their ability to adopt and use it successfully?
b. How does the training on using the new AI system experts receive affect their ability to adopt and use it successfully?
c. What factors interfere with experts' ability to use AI systems when they first try to use them independently?
d. What factors interfere with experts' ability to use AI systems after they have already used them for a while?

3. **Outline the perceptions of the reasons for low in-house AI software adoption among experts, AI developers, and team leaders within the science organisation.**

a. What do experts see as the key factors influencing their willingness to adopt a new AI software?
b. What do experts see as the key factors influencing their choice to reject a new AI software or search for a replacement?
c. What do AI developers see as the key factors influencing experts' willingness to adopt new AI software?
d. What do experts see as the key factors influencing their choice to reject a new AI software or search for a replacement?
e. What do team leaders see as the key factors influencing experts' willingness to adopt a new AI software?

f. What do team leaders see as the key factors influencing experts' choice to reject a new AI software or search for a replacement?

4. **Identify how different stakeholders' perceptions align or misalign and how that influences the adoption of a new AI system.**

g. How do these perceptions differ between experts, team leaders and software engineers?
h. How do these differences affect the strategies team leaders use regarding AI adoption, including planning introduction, support and training?
i. How do these differences affect the way software engineers interact with experts regarding the development of in-house AI systems?
j. How do these differences affect the way experts interact with software engineers?

5. **Understand how AI developers, team leaders and experts approach the issue of low AI adoption.**

a. What strategies do experts use to improve their ability to use in-house AI software more effectively?
b. What strategies do team leaders use to increase experts' adoption of in-house software within and outside of their organisations?
c. What strategies do software engineers use to increase in-house software adoption by experts within and outside of their organisations?

6. **Explore the role of collaboration between stakeholders for the AI adoption**

a. How do experts interact with team leaders and software engineers?
b. What effect do the presence and absence of these interactions have on experts' ability and willingness to adopt in-house AI tools?

4.3.5 Participants

Ten participants took part in the study. One AI team lead and science director and four AI software engineers were interviewed. Five scientists were observed and interviewed (Table 1).

Table 1: The list of participants

Participant	Participant group
P01	AI software developer
P02	AI team lead and science director
E03	Scientist
P04	AI software developer
E05	Scientist
E06	Scientist
E07	Scientist
P08	AI software developer
P09	AI software developer
E10	Scientist

The AI software developers and expert participants were recruited using snowball sampling (Goodman, 1961). First, the AI team lead and science director informed their team members about the study and shared the invitation to participate. The AI team lead and science director also informed the other team leaders within the organisation, who then shared the study invitation with the experts in their teams. The inclusion criteria were participants working in the office and being available for in-person interviews and observations. For expert participants, using AI-enabled software had to be part of their project(s). Participants voluntarily participated and were not compensated for their participation due to the internal rules of the organisation in which the research was conducted. All participants were sent a digital copy of the consent form and information sheet to sign via email at least a day before their participation. They were asked to return it to the principal researcher before the observation/interview. See Appendix A for the study consent form and information sheet.

4.3.6 Procedure

The first interview with the AI team lead and science director lasted around 90 minutes and was conducted in person in their office. The participant signed the consent form before the interview and was briefed about the study aims and their right to ask questions and withdraw at any time. The team lead and science director

demonstrated additional information during the interview, such as the training booklet.

Interviews with software developers lasted around 45 minutes and followed a similar procedure. They returned a signed consent form, were informed about the study and their rights, and then asked interview questions. Software developers were encouraged to demonstrate any aspects of the software they saw as important for supporting their answers. The interviews were scheduled during their working hours and conducted in pre-booked conference rooms at the organisation.

The observations and interviews with expert scientists lasted around 90 minutes. They were conducted during their working hours in the lab, where experts usually used the software or a pre-booked conference room. Experts were first asked to answer general questions, then demonstrated the software they used and answered observation-related questions throughout. Following the observation, the semi-structured interviews were conducted.

4.3.7 Interview protocol

First, all participants were asked general questions about their roles, practices, and workflows to gain background information and to understand their work context. All interviews were semi-structured and guided by participants' answers. The AI team lead and science director was asked questions about their team, the expert team, the development and implementation of their in-house AI software, training and collaboration between teams. Software developers were asked how they developed and introduced usable software to potential users and how they developed and conducted training sessions. The third part of the interview was focused on collaboration experiences. They were asked to share experiences and expectations regarding collaboration between them and experts. Expert participants were asked how they chose and trusted an in-house AI software, how they felt they understood it, how they were first introduced to it, and what training they underwent. They were also asked about their experience with and attitudes toward the in-house AI software they used. Lastly, they answered questions about their collaboration with software developers. The full interview protocol, including the interview questions, can be found in Appendix B.

4.3.8 Data collection

Observations and semi-structured interviews were audio recorded. Reflective notes were taken during the observations and after the interviews. The interviews were transcribed using Trint software (https://trint.com/), going back to the actual recording to check for clarity and correctness. Collected notes were only used to inform the data analysis and write up the results. The parts of the transcriptions that were used were edited to remove personally identifiable and confidential information. All participants were informed about their right to withdraw from the study at any time and for any reason and have their data removed from the study with no penalty or loss of benefits to which they may be otherwise entitled. The Edinburgh College of Art Ethics Board approved the study. More details about the study data handling, access, storage and privacy can be found in Appendix C.

4.3.9 Data analysis

Data was analysed using the Grounded Theory Approach and Advanced Coding Method (Chun Tie et al., 2019). Initial codes were used to divide the data into categories, which were then transformed into more abstract concepts and their dimensions. The relationships between categories and the central core category were identified, creating a storyline supported by explanatory connecting statements (Corbin & Strauss, 1990).

4.3.10 Positionality

I spent four days working in an open office alongside the AI team members and science experts, where I was part of informal conversations and observations that contextually enriched the collected data. Because of the rich contextual understanding of the data that was gained while visiting the organisation and collecting the data, I also conducted the data analysis. I acknowledge that their experiences and positionality could have influenced the perspective and approaches regarding data collection and analysis.

4.3.11 Limitations

The proper contextual inquiry study (observing users doing their usual work) was not possible. While some experts brought their data to analyse using the in-house AI software, others did not, as they did not know how to use it, or it did not work. Even

the experts who brought their data to analyse spent most of the time trying to get the software running. As a result, the observations were mostly of them trying to open the software and demonstrate how they would use it rather than of them conducting their actual work. However, these observations were still valuable contributions, showing how experts did and did not interact with AI software at work.

4.4 Results

This section covers the key categories that emerged during the analysis and the interrelations between these categories and a core category 'assumptions'. The answers of the practitioners and experts are juxtaposed to emphasise the miscommunications between the teams and link them to the key frustrations. The letter "P" next to the participant number indicates practitioners (software developers), including the AI team lead, and "E" refers to the science experts. In the interviews, experts referred to AI-enabled systems as software, and this term is used throughout the following sections.

4.4.1 Software development and user training

This part outlines the barriers present before the software implementation and shows how they can create friction in the later AI implementation stages. Each barrier is viewed from both experts' and practitioners' perspectives.

4.4.1.1 Barrier 1: Users' knowledge of AI/ML

When experts were asked about their general understanding of AI and ML, most of them admitted that they lacked it: *"My understanding of machine learning is probably disappointingly low […] it is at the level that I can use the tools […] but I do not have a very deep mechanistic understanding."* (E03) Interestingly, the expert was able to successfully demonstrate and explain different software (not the in-house one, because it did not open), suggesting that they might be underestimating their understanding of AI and ML. Other experts said that they lacked computational knowledge but were able to learn with the proper support (this expert was collaborating with the development team): *"Not very high. Before I came here, I could not even code. And then in the first few months of our program […] we got some basic introduction to Python."* (E10) One of the experts said that they were able to use the software without having extensive programming expertise: *"[…] I do not have*

high-level programming knowledge. I will at the end of my [project], but not right now, but I can use it. No problem." (E05) This expert was closely collaborating with the development team, and they had some programming experience prior their project. On the other hand, an expert who was not in the collaborative project expressed the difficulty of learning about ML on their own: *"[…] I have been trying to keep myself sort of in the loop a bit more about how these machine learning mechanisms work, and I would not say it has been very successful so far."* (E06) The same expert said that just by using the software without support, they were only learning to use the software, but not gaining an understanding of ML: *"[…] I would not say I am understanding more about how the machine learning works. I am just understanding how to use the software."* (E06) Another expert admitted that it has been challenging to learn to use the software: *"It has been a big learning curve in the last year." (E03)*

4.4.1.2 Practitioners' assumption: If they cannot use it, they should not use it

The AI team lead and science director referred to the minimal computational knowledge requirement as a safeguard for the technology not to be misused by experts. The assumption is that if the user does not have a certain level of understanding, they should not be using the tool at all. Adding a certain difficulty to using the software (or not making it too easy to use) is seen as a barrier: *"I do think that with some of our techniques, you need to know a little about what is going on. Otherwise, they can give you horrible, horrible results. And you were just unaware of the fact that you are getting a horrible result. […] I think the safeguards are that they do not use it [software] because they do not know how it works."* (P02) They said that to prevent the automation bias, they introduce the knowledge barrier of entry: *"[…] I think being wary at that level of trust, we are introducing barriers to our users […] if we were to reduce the barrier to entry by reducing the amount we tell them about how the tool works, the tool becomes much more of a black box."* (P02) However, there is no strategy to help experts who cannot overcome this barrier. *"I think ultimately, we do want it to be a tool that people can use […] I guess you need to put it in a state where the barrier is not so high that nobody enters, but that the barrier is high, but the benefits are obvious and that they can see, yeah, okay, well, I understand why it is worth me investing this time in upskilling my skills into understanding this new area because it will really benefit."* (P02). The interviews with science experts later showed that they were often unaware of the benefits of the

software, as they were not communicated to them clearly and understandably. One of the software engineers said that experts lacked understanding of AI and ML because they were not interested in learning about them: "*[scientists] are not so much interested in computing. They just want the end result. They just want something that will work because they have so many other things that they want to be thinking about. […] We are expected to learn the biology, but they are never expected to learn the computing side of things*." (P08)

Observing experts and listening to them, it became apparent that they had a high-level understanding of how AI and ML worked and had basic computational knowledge. Moreover, all the interviewed scientists had high expertise in quantitative research methods and statistical data analysis. This suggests that experts might not be unqualified to analyse the system outputs. However, they lacked some basic programming skills necessary for running these systems that practitioners assumed were "common sense". As a result, experts did not receive the necessary support, and the software was unattainable for them. Introducing the computational knowledge threshold without providing the necessary support could disadvantage experienced scientists who need to use the software for their research.

4.4.1.3 Barrier 2: Assumptions about users' training needs

During the software training, experts have an opportunity to try it for the first time, discover its features and learn how to use it. The training workshops are half-day or full-day events intended to teach potential (or existing) users about how the software works, how they can use it, and what its limitations and capabilities are. However, experts referred to the training as challenging to follow because the teaching staff were explaining information from their perspective: *"[…] I have had sort of tutorials with some AI team people about using [software], and we have tried to use it on our datasets. But I think it is very challenging for people who think about things very differently to teach each other about such because I am obviously so focused on the biological question of my data and things like that, whereas they are focused on, you know, the machine learning mechanisms […] I think it can be very challenging for someone like that to teach someone like me how to use AI software."* (E06) The same expert said they found the training confusing: *"I was just getting quite confused about why we were doing certain elements of the workflow and what those different elements were actually achieving with, you know, that the whole idea of what we*

were trying to achieve by segmenting the whole dataset. I did not really feel like I understood why each step of the process was necessary." (E06) Another expert said that the amount of effort that training required made them think that using the software would be too much work: *"I felt like that was quite an extensive training requirement […] I have not really used it [after the training] for what I was trying to do here. It just seemed more work."* (E03) Given that all the interviewed scientists had this training but could not effectively use or open the software, it suggests that the training goals did not translate into practice.

4.4.1.4 Practitioners' assumption: Training is just training…it is not a big deal

When I asked practitioners how they decided what information should be included in the training, one of them said that they tried to avoid going into much technical detail: *"You have to explain them [experts] sufficiently, not to get them deep into maths […] in an abstract way that they will understand it, but not giving them a lot of details that it would be like the university module."* (P01) The expert also said: *"[…] we try to also explain how exactly it works to the level they need to understand […] what are the trade-offs, but not get into how exactly the linear algebra works."* (P01) The AI team lead referred to a certain level of understanding of the underlying software mechanisms needed to use the software effectively: *"[…] if you know how [the software] is working behind the scenes, you can use it in the most effective way. But that said, what it does mean is we do go into what it is doing reasonably in detail, not hugely."* (P02)

However, when asked how they assessed the effectiveness of the training, practitioners saw the success of the training as users' responsibility. The AI team lead said: *"If people follow the instructions reasonably closely, they get quite good [results] out of some quite tricky data, and hopefully that gives them the reason we wanted that to happen was so that at the end of it they get the idea of, okay, well I can see how this works for this."* (P02). This suggests that if they did not understand it, it was because they did not follow the workshop attentively. However, when I asked how they tracked whether users applied the knowledge after the training and whether they used the software, the AI team lead admitted that they had not done much about it: *"[…] I do not think we have a particularly great conversion rate […] What happens at the moment is that we generally run like the exit questionnaire ourselves, and people say, yes, it was fine, and we ourselves on the back and*

congratulate ourselves for another training course well delivered […] I do not think we possibly reflect enough on actually the impact that that is had […] And I think if we really evaluated the output of those." (P02) This emphasises the lack of feedback-based ongoing communication, reaching out to experts for more in-depth feedback about its quality or trying to design and edit its material by collaborating with experts. The AI team lead also admitted that training has been more of a formality, and they did not think about it much before this interview. *"You need to provide the training materials, but then that is it […] it is not a huge thing."* (P02) Then they said: *"[…] we kind of treat training as something we need to do, and I wonder if it is something we need to actually put a bit more effort into specifically around how to get the best outcomes, how to make sure people have actually learned because, you know, we do not validate anything […]."* (P02) This question led to a deeper reflection about the value of current training practices: *"I wonder if actually what has happened is that we have kind of sort of dropped into the this is something we have to do because we said we do it. […] So, we put together the course, we run the training. We do not really think about it long term in terms of any of the benefits."* (P02)

4.4.1.5 Barrier 3. Lack of information about the software capabilities

Experts needed to know the software's capabilities before investing time in it. *"It is difficult when you do not even know if the software has the capabilities that you need it to have. I do not want to put a lot of time into learning how to use something when I do not actually know if the software itself is good enough for what I need it to be able to do because that would just be a complete waste of my time."* (E06) They were willing to learn more about the software if they saw it would be useful for their project. *"[I] am purely learning on the job, but when you know that you want to use it, I find it interesting, and I know it is going to give me what I want."* (E10) When not knowing the functionalities of the software, experts were more likely to seek information from other experts and online sources, such as Twitter. *"Having lots of feedback from other users is like the main metric that I would use as to whether the software will be useful for me or not […] but a lot of that again just comes from Twitter, which is kind of ridiculous, because unless you are on Twitter, why would you know whether a software is good or not, it seems?"* (E06) They found that the information online was especially helpful if it used examples similar to their use case*:*

“On the website, they were like […] here is an example of someone using this software for this specific dataset or this specific dataset, you could at least get an idea of whether it might be applicable to yours.”(P06) This shows how important it is to communicate software qualities in a way that is relatable to experts and their specific tasks. If the software is new or not well-presented online, the users are more likely to choose a different software that they at least know could produce something useful for their research.

4.4.1.6 Practitioners' assumption: They will eventually find out

One of the practitioners admitted that software capabilities were not communicated in a way that was relatable to the experts: “*The scientists did not know what [software] could do. It was not explained very well because the way that it was presented in our science talks and stuff was very much like a computing thing […] the presentations will always like what does the [software] do rather than what does [software] do for you […] that is, I think, why uptake was really low."* (P08). The same practitioner shared a success story of applying this practice: *“We took some data which people were actually working with and showed them what they could do, and they were like, actually, this is really cool, why did we not know about this before? And we were like, it has always been here. It just did not get across.”* (P08) The AI team leader explained that they waited for a journal publication explaining what the software could do*: “I have been waiting for a big, nice publication that says, well look, here is how this works, here is how you use this”.* (P02) They also expected experts to independently recognise the benefits of the software and advocate for it to their peers: *“It is a case of you need to find somebody who picks up the tool and goes, actually, yes, this works.”* (P02). They also thought that experts often had unfound expectations for the software: *“I think there is also a slight misconception that AI will solve all your problems and it will solve them very easily”.* (P02) However, these misperceptions might suggest that experts were not effectively communicated the actual capabilities of the software. When AI team lead was asked why they thought the experts did not adopt their software, they listed factors such as the competitors’ software and experts using what has been available for a while rather than trying something new*: “[…] there are already other things in the field that people are used to using that do not necessarily work as well as what you have got, but they are what people used to do.”* (P02) This suggests that the AI team see experts

rejecting their developed software due to the fear of change and resistance to trying something new. When talking about the software adoption challenges, practitioners did not mention that experts might not know how it could be useful to them or whether it would perform better than the software or methods they were used to use.

4.4.2 Using unfamiliar software

Interviews and observations revealed challenges that experts faced when starting to use new AI software. Here, I detail four key frictions that occurred during the initial stages of software use. Each friction is juxtaposed with the practitioners' associated assumptions to illustrate how misapprehensions between them could lead to unresolved issues.

4.4.2.1 Barrier 1: Not knowing how to get it to work

When experts were asked to demonstrate the software they had to use, three out of five could not open it (these were different systems). All observed experts struggled with the command line and had to search for guidance in the available documentation online. They were frustrated and expressed how stressful the installation and initiation processes were without having the programming knowledge: *"[…] it is absolute hell to try to install in the first place. And if you do not have any programming background, you lost."* (E05). Other experts said: *"[…] I find the command line the most difficult thing ever."* (E06). When asked how they know the command line, the experts said: *"Hopefully, someone has told me at some point, and I will have remembered it, but that is not always the case."* (E06). They also referred to the lack of appreciation of their background: *"[…]I feel quite debilitated by the fact that I'm a biologist, and that is what I understand. And then I came to this software and tried to use it, and I feel like it is very much built for people who have a much better understanding of that kind of area of things."* (E06)

Despite AI tools being intended to increase the productivity and efficiency of its users, interviewed experts admitted that trying to get the system to work was time-consuming and took their focus away from their projects: *"I have only started my project […] and a lot of it has been trying to get something to work, installation and stuff like that."* (E05) The ones that were still looking for the right software were fearful about how long it would take to learn how to use it: *"[…] this is like a major stressor within my project […] I am just going to have to spend ages sifting through*

this different software, trying to get them to work without having a professional with me who knows how to work it all out. So, it is just going to be painful." (E06) They added: *"[…] it is completely taking me outside of what I actually want to achieve with my project because I am having to put so much time and effort into it."* (E06). Another expert felt the expectations for them were too high and developers did not take into account the amount of work experts already did: *"We all have to be experts in quite a few different things right […] we cannot, we do not spend all of our time just focusing on the processing. There are lots of different things that we have to be doing […] so yeah, we are primarily not programmers […] Obviously, it is easier the more computer literate you are, but it is always an unreasonable expectation that we are all going to be, you know, pro computer people. I think it is very important that the software is accessible for fairly standard users."* (E03)

When asked if they had spent more time dealing with software issues than working on their projects, one of the experts said: *"Yes. Which I have heard is quite normal, especially for scientific software because it is all very specific."* (E05) Not being able to start using the software stifled experts' research progress: *"I just have loads of data that is sitting there that have not been segmented, has a lot of biological information within it, but I have not been able to take any of it. The only way I can see you doing that at the moment is to start doing manual [analysis], which would obviously take me a lifetime because there are hundreds of datasets."* (E06)

4.4.2.2 Practitioners' assumption: "They must know this much"

Practitioners admitted that experts need to know fundamental aspects like the command line. Still, it should not be difficult: *"You probably need some basic knowledge about Linux and command line and manipulation, and you do not need any really advanced programming knowledge […] there is a lot of manual clicking but not actual programming involved. So, it is quite easy to use" (P04). Another practitioner said: "[…] You need to be able to use a command line, and you need to be able to think about image data, which often has various sort of very specific qualities relating to the quality of the image or the resolution."* (P09) Practitioners were aware that experts might lack the necessary skills. Still, they overestimated them: *"You assume that these days a lot of people have some knowledge of machine learning if they are a scientist […] you assume a certain background, but you are probably often not right."* (P09) It was not seen as an issue by this

practitioner: *"[…] some people with less programming experience have been able to run it. A lot of the people who use [software] do have some programming experience."* (P09). Another practitioner expressed surprise at the lack of experts' technical savviness: *"I have worked with somebody recently, who was effectively a user […] and the level of computational knowledge that they had was incredibly limited to the point of things that I do not even consider. It is not even something you think about anymore."* (P02) This suggests that some aspects of running the software were not even considered by practitioners. They saw them not as easy or difficult but as common sense.

Moreover, tasks that were not directly related to software development (i.e., help with software installation or updating instructions) were seen by practitioners as secondary, something that someone else would do: "*Somebody in a group [scientist team] that wants to use [the software] has the technical expertise and maybe does the job of installing the software, and then other people in the team with less technical expertise use it."* (P09) The same practitioner said that in the training sessions, they installed the software for users to be ready to run the training tasks on: "*[...] in the workshops, what we do is we already have it installed […] we made it pretty straightforward so people could type in, like we had a card of information which had a name and a password, and they were able to type that in on their laptop and the [software] was ready to run."* (P09) This suggests that experts were never actually trained to do that before starting to use the software.

4.4.2.3 Barrer 2: Incomprehensible, outdated, or missing documentation

Experts emphasised the need for easy-to-follow and up-to-date installation instructions that would not require them to have a certain level of programming knowledge: *"Very clear installation instructions like almost a step by step […] then also step by step what code you should be running and an explanation as to what each of the bits of the code is so that you not have to change it for your specific work. Um, because it should be that anyone can use these. You should not have to have years of computer science background."* (E05) Another expert emphasised the importance of comprehensible and up-to-date documentation: *"Easy to install, easy to understand […] having good documentation is probably the main thing."* (E03) Another expert mentioned how important it was that instructions are prepared considering their level of computation knowledge. *"So many of these softwares are*

just absolutely horrendous to use. […] A lot of the time, there are no instructions online on how to use it, or if they do, it is for people that are very specifically computer science." (E05). An expert who has already learned to use AI tools admitted that it had been difficult: *"The thing I found challenging at the start was that I never used Linux before. And here is all Linux. I think that was one of the hardest things I found, just understanding how to use the system and just so many different pieces of software to use. And the documentation, as I said before, there is no established [documentation].*" (E03)

4.4.2.4 Practitioners' assumption: It is available, so it must be useful

Practitioners explained that documentation is by default made available to accompany new software: *"[…] we always try to make all our software sustainable and to have good documentation […] so they can visit the documentation at their own time get into deeper end on how to use the software."* (P01) When explaining how users would know how to get a new software working, they explained: *"[…] it has a GitHub page, so it has a repository, and it also has, I think, online documentation, which is quite self-contained, just good."* (P04) When talking about the software they developed, a practitioner said, *"There is the documentation, the manual, and then there is putting an issue on GitHub or sending us an email there was a forum that we had for a while."* (P09) But after being asked if there was anything apart the paper booklet for one of their software, the same practitioner said: *"Yeah, it is just that booklet, that kind of thing […] it is likely to be important for them to be able to run [software], especially to install."* (P09) The practitioner referred to the physical guide booklet that one of the interviewed experts tried to use when they could not run the software: *"Luckily, I have remembered that somewhere there's this very handy guide. It is handy because it does explain some levels of how you use [the software]. Whether it tells me how to open it is another thing. Sometimes, I find this guide is just too simplistic. It leaves out a lot of information that I need."* (E06). The expert did not find the necessary information in the booklet. However, it also appeared that experts had not been contacting practitioners with questions about the software and related documentation. As a result, practitioners did not see documentation as a factor that could influence the adoption of the software and did not seek to investigate whether users were happy with the quality of their provided software documentation. One practitioner associated no complaints as a positive

indication: *“We have not had too many problems or people seem to be able to learn.”* (P09)

4.4.2.5 Barrier 3. Lack of active support when starting to use the software

For experts, asking for help from the in-house AI team was the last option. When asked what they would do if something did not work, they said: *“Ask someone else here [among colleagues] first. There are good forums online […] I have been working with a postdoc here, and he has basically taught me and all of this pipeline.”* (P03) Another expert said: *“I could give you the easy answer to it - I struggle. […] I think most of the time I try to find material that is out there that maybe other people have made to try and help the users to use the software and things like videos online.”* (E06) This expert did not have anyone to contact and had to rely on the Internet to find answers: *“[…] I do not have anyone more senior to be able to turn to, to be like, Hey, I see that you have used this software before. Can you help me learn how to use it? […] I am just going to have to work it out by myself. I am just trying to look on the internet and see if there are any tutorials on there and things like that, which is obviously challenging”* (E06). The expert was emotional about not having anyone actively supporting them at the start:

> *“I think it is most frustrating because I know that if someone was around to just teach me how to use it, we could then spend some time, sit down, work it out, practice with using it on some of my data sets. It would probably take a day or two, and then it would simplify the rest of my project. Because then that problem would just be solved. But because those expertises are not here, or if they are, they are not lending themselves to me, it just means that I'm having to struggle through trying to get something to work, which makes you feel stupid sometimes because you are trying to get the software to work and you feel like, Well, why I cannot get this to work? You know what is wrong with me: I cannot get this to work. It can cause some real existential crises sometimes.”* (E06)

Another expert also emphasised how challenging it was to find solutions independently: “[…] I was trying to go through literature as to how maybe people who have helped develop it a little bit were using it to try and figure out what code I was meant to be using. Um, no joy whatsoever.” (E05) One of the experts explained that

it was intimidating for a scientist to admit that they did not know something basic, so they did not reach out to the AI team with these questions: “There is definitely an element of pride […] I would say there is an element of not wanting to admit that you do not understand something.” (E06)

When experts worked together with practitioners’ support on seemingly straightforward tasks, it improved how they approached the software and how they dealt with its issues later: *“I kind of worked with them on learning how to use the software [...] which is quite straightforward software, but when you're starting, and you have no idea, it is useful.”* (E10). Providing active support at the beginning could save practitioners’ time in the future: *“[Practitioner] was trying to get involved in actually seeing how the code runs and everything and seeing how we were able to run it ourselves eventually. So [practitioner] does not have to every time we want to use something run it for us.”* (E10) Getting answers to simple questions led to learning about the software and being more confident and independent in using it. This also meant they did not have to ask these simple questions every time they used the software. *“[practitioner] has been fantastic at not just telling me what to do but helping me to understand a bit deeper as to what I am supposed to do in the future […] he is helping me so that I am not as lost every single time I use it.”* (P05)

4.4.2.6 Practitioners’ assumption: They do not want to put work into learning about the software

When asked about troubleshooting, one of the experts said: *“I have taught programming for quite a few years. We always tell students, go see your friend, search if it is open source, if it has a repository and GitHub, and then if they are well developed and well looked after, then they should have a link to the online documentation.”* (P04). This answer also shows that the practitioner compared experts to their programming students, showing a certain level of expectation they have for the experts, assuming that the same type of material and training methods should be suitable for both groups. When they did have to provide support, frustration was evident: *“I just sent them a very explanatory email of how exactly you have to do it, what you should do to use it […] what is the correct word about it, customer service or something.”* (P01) Practitioners felt frustrated by the way they were being approached by the experts: *“[…] most of the time they are like, I do not know how to do this, please fix it. And there is no clear like, please, can you do*

ABC? It is up to us to figure out ABC and then we work with them again to say well, I manage to do ABC for you. What do you think? And like I actually wanted X, Y, Z. So, there is sometimes a mismatch in expectations from both sides." (P08)
After introducing software, practitioners did not reach out to experts to find out whether they were able to use it: *"We train them to use the software and then let them get on with it and then the process of working through. You know, it is just a case of emails that backwards and forwards, if they have questions."* (P02)
Practitioner admitted that they expected experts to contact them: *"No, I do not do it actively. It is up to them."* (P09) The AI team lead also said that experts simply did not think they would have to put any work in, and when they saw that it still required their effort, they rejected the software: *"We do a lot of work on this tool and because they see us actively working, they assume, okay, well, this is going to work for us. And then when we say to them […], you are actually going to need to still put in a few hours work. You are still going to have to do this. It is like, well, okay, but I was hoping for something that would just give me the answer. So why is this not just giving me the answer?"* (P02)

4.4.3 Effective collaboration and communication as a solution

This part reflects on the collaboration between experts and practitioners. Here, I present cases where collaboration led to removing some of the frictions described above and outline factors that led to successful collaboration or hindered it. Two interviewed experts worked with the AI team as part of their projects and were more positive and involved than other experts. Software developers reflected on their experiences of collaborations as successes. These stories involved users becoming advocates for the software, getting involved and finding the confidence to approach each other.

4.4.3.1 Collaboration can help establish mutual understanding over time

In one project, a group of practitioners and experts had to collaborate on developing the software in weekly meetings. The communication between the teams was not effective in the beginning: *"At the start, we could not communicate with each other at all."* (E10) The expert said that they simply did not understand each other's languages and had to have someone with both backgrounds to translate. *"We were speaking in chemistry, and they were speaking in AI. And I genuinely felt like we*

were speaking different languages, and the meetings would go round in circles a lot." (E10) After around three months, they became better at communicating. Having a shared goal helped: *"But we have gotten a lot better at it. […] We needed time where we all learned a bit more to be able to talk to each other. I think everyone would agree because you are trying to figure out how to make [software] better."* (E10) One practitioner also reflected that working with experts helped to understand scientists better: *"I find it easier to talk to a biologist now that I have spent a lot of time with them […] But if you are someone who is completely in your field and all the people you speak to on a daily basis are computer scientists, if you are trying to speak to a biologist, it is a bit difficult to bridge that gap."* (P08)

4.4.3.2 Continuous collaboration can help to align expectations

Collaborating appeared essential for effectively aligning expectations and finding middle-ground solutions: *"Mostly what those meetings consist of is all of us talking, right? Does that make sense as a chemist? […] From my point of view, it is a lot of just kind of hearing what [practitioner] trying and try to make sense of whether that makes sense from the chemistry side of things."* (E10) Effective collaboration was based on maintaining the relationship throughout the software development process and frequently checking in with each other: *"There is this continuous communication regarding how the computer science attempts to solve certain issues and if this is okay with them [experts]. For example, for what I am doing right now, we initially started with idea number x0, and now we are at x5. It evolved around what we can do and what we cannot do. I propose a way to do it. They may counter-propose that this may create other problems elsewhere. Then there is a little back and forth to find a middle ground and something that can be done."* (P01) Continuous communication led practitioners to reflect on their assumptions about users' needs: *"[…] now that we are trying it with them, sitting with them, we are finding that they do things differently than how we expected them to."* (P08) It also helped them think about the usability of the software: *"[…] I refused to write a GUI (graphical user interface), and the code was initially intended for use locally. But the scientists said that there is no GUI - not using it yet. So that is why that sort of prompted me to think about, yeah, it is not only the developer's content, and it is for the user." (P04)*

The AI lead also emphasised the importance of communicating throughout the development process: *"It is very easy from a software development point of view to*

write a piece of software that is not usable or does not do what you want it to do. And so having that constant engagement with the researcher." (P02) One of the practitioners said that experts' knowledge gave insights into their research field and helped to consider details only experts knew: *"It is very important […] to bring in somebody who is an expert and get some feedback from them because otherwise, you can spend a lot of time just talking about blobs, and you are not talking about what you need to be talking about."* (P09) Another practitioner added that practitioners should understand the needs of experts to develop useful software: *"I think it is really important to spend time with people in the domain that you are working with and understand what is important to them before we can help them do anything."* (P08) The practitioner admitted that waiting until the software is developed and then talking about it with experts was not practical: *"[..] So [one of the practitioners] went ahead and spent months writing this thing. And then because he was in his little silo, writing his own thing and the scientists were busy doing other stuff […] when [practitioner] finally was able to give them a solution, the first version they were like, I actually wanted it to do this, but three months ago. Now, I wanted to do a completely different thing. And so that is very frustrating because [practitioner] spent so long doing it." (P08)*

4.4.3.3 Collaboration increases experts' motivation to be more involved

When experts had to collaborate with domain experts due to their project requirements, they actively provided feedback about the software to the AI team. The AI team lead reflected on one of the success stories: *"This user was basically happy to sit and use all the software that did not work […] Okay, well, this does not work. I cannot do this, and this is what I really need today. That enabled us to really drive the software towards something that was practically useful."* (P02) One of the collaborating experts said that they were independently trying to improve the software and persisted in using it even when it performed poorly: *"I run two [analysis] and they were not fantastic […] [AI team] are kind of working on figuring that out at the moment. So, instead of just sitting and doing nothing, I am going to try and see if I can get my [inputs] a bit more specific […]." (E10)*

During the collaboration, experts also had a chance to learn about software capabilities and communicate their needs to the practitioners. It also allowed them to see the benefits of even imperfect software and be willing to advocate it to other

potential users. When the AI lead was talking about one of the experts who was using and advocating for their developed software, they realised that it was because of the close team collaboration: *"Now thinking about it, he is one of the people that we think would be happier to use the software because he has experienced it. But that is because he used it with the developer, and it was collaborative. It was not [because of] using the tool. It was a collaborative approach."* (P02) One of the experts who collaborated with developers was willing to invest time and effort in trying the software and helping to improve it: *"Having played around with it and seeing like what the problems I experience are, it is quite nice coming up with ideas, being like, okay, if I fix this, this is going to make this so much easier for everyone."* (E05) The close collaboration encouraged this expert to be actively involved in testing new versions of the software as they saw the value of their input: "[…] *you are kind of helping beta testers, if they are kind of like ideas you have, um, if you kind of monitor, like, how long it takes for you to do it, how well each the individual bits are working. If you can kind of make a note of that, and then we can kind of discuss it further with, obviously, the groups that develop it […] I guess me not having a programming background helps on the whole kind of like, how is the average scientist going to use this?" (E05)*

4.4.3.4 Collaboration makes communication more casual

One of the experts shared how the courage to ask questions increased while collaborating with the AI team. *"If I do not get it, I will just be like, what are you talking about? Whereas at the start, I was like, oh, they expect me to know this. I think it is more of a personal thing. […] I was like, I do not want to waste everyone's time here trying to explain to me. […] I know now they would not have thought it was a waste of time in my mind."* (E10) The expert shared that constantly talking to practitioners made it easy to approach them: *"[…] we talk so much, we spend so much time together, and I would just be like, yeah, hi, remember me? Um, I have this problem."* (E10) An expert who already had an established connection with practitioners was happy to contact them: *"I would normally email the software developers and be like, yey, this is for some reason not working or isn't working as well as it should."* (E05) They were also happy to give feedback about errors and persist with the choice of the software: "*I would much rather contact the development team, mostly because I know them now."* (E05) Without having an established connection, it was also

difficult to give constructive feedback unless it was encouraged and/or initiated by the AI team: *"[…] it is just always difficult to try and give out feedback in a sort of constructive way". (E06)* The same expert admitted that they saw the benefit in providing feedback: *"But I do understand that if I give them some feedback. It might help the software to be better for others as well […] I think, you know, especially when you're using this kind of software for answering some biological questions, having that biological context will be important and to be able to input your knowledge of maybe what you should be saying, or you would expect to see back into the segmentation is going to be useful." (E06)*

4.4.4 Trust and explainability

This part presents the experts' and software developers' views on trust and explainability. The answers illustrate the key point of this chapter – explainability can only be effective when users' fundamental needs of training, initial support and collaboration are met. In the study, experts did not even think about trust and explainability because they focused on how to get the software working and then how to use it and troubleshoot it independently.

4.4.4.1 Trust judgements

When asked how they decided whether to trust the software and its outputs, one of the experts admitted that they had not thought about it before they were asked the question: *"Actually, I do not know that. That is a very good point. I suppose I just look at what it has done and try to assess whether it seems like it has done a good job or not, according to what my credentials are for what a good job is."* (E03) However, the expert admitted that trust issues were going to be important once they started using the software regularly: "*[…] that is a good point because obviously going forward, once I start using the software more habitually looking at my data, then I need to have a metric for knowing whether the [process] that it is done is actually correct and whether I can trust the data that it is given me. Because yeah, it is a very good point."* (E03)

When asked to speculate how they would decide whether they should rely on the software, the expert said that established tests and others using it would indicate that it was trustworthy: "*I guess once you have thorough enough testing and you know, enough people start to use it, and they are getting a reasonable result out of it that*

are also validated with a more traditional pipeline." (E03) However, after a short pause, the same expert said that they would rely on their expert knowledge: "*Gut feeling, exactly my gut feeling. Yeah. I mean, I suppose I have the contextual understanding of the data that I am looking at to be like. […] I guess I would probably trust my own judgment more."* (E03) Another expert suggested that having metrics indicating how reliable the software or its outputs were could be helpful: "*I suppose metrics of some kind are always going to be useful, to have percentages of confidence and things like that. I cannot really think of any other way of knowing how much you trust it*." (E06)

4.4.4.2 Manual assessments

The expert who used non-in-house software said that they had spent a lot of time manually inspecting its outputs, as they still did not trust the accuracy of the results. When I asked how they decided which outputs they would check manually, they said: *"I would go back and check a few of them […] And if a few of them look alright, then I would probably trust it. But if in those few I saw several points where it looked like it was getting it wrong, then I would go back and check more of them and then probably kind of readjust the thresholding or maybe consider doing it manually or using a different software […]."* (E03) When using a different piece of software, the same experts demonstrated how they would manually inspect it. When asked about the validation feature that I noticed on the screen, the expert admitted that they preferred manually checking the results: *"I personally normally prefer to just look at, you know, I said you actually manually go back to look at it visually. That is normally how I prefer to do it."* (E03) When I asked why, the expert expressed distrust in the software: *"I guess. I think even if I did look at this validation threshold, I would still, I think it is easier to interpret visually if you know what you are looking for."* (E03) When I asked if they would trust fully automated systems, the expert said: *"Possibly. But I think doing something like manually checking, that really does not take that long […]."* (E03) These responses suggests that simply seeing predictions, even when they are explainable might not be satisfactory to experts and they would still want to use their expertise to manually double check things.

4.4.4.3 Explainability

Experts had not used explainability features or thought of explainability before the interviews, but when asked if it would increase their trust in the software, one of the

experts said that it would increase transparency and understanding, increasing the trust as a result: *“Yeah, I think so. I think it would add more transparency, which normally means you trust it more. I think that the more you understand it, I guess that is what it comes down to. The more you understand something, the more likely you are to trust it. Yeah. And where I do not necessarily understand it, then I am more likely to go back and to manually check it myself.”* (E03) Another expert said that explainability would probably just add more complexity and could only be useful if it was simplified to their level of understanding: *“Maybe if it was a very simplified explanation, but because like, obviously if it is going to be like, oh, we used, you know, this kind of nets to do this, I am going to be like, okay, you have lost me. I do not understand what you are saying. I think it would have to be a very simplified way of explaining what it had done.”* (E06) This aligns with previous chapters showing that explainability is often not seen as useful by experts and can be rejected even before properly trialling it if it does not appear designed for experts. When asked about the importance of understanding the underlying AI/ML structures and processes, experts said it was not necessary: *“[…] I do not really need to understand all of the architecture behind it. That's kind of, at least for me, how my impression is.”* (E03) Another expert said: *“[…] if I do not need to, to be honest, I do not. Because it is not what I am interested in, you know, I'm interested in biology.”* (E06)

4.4.4.4 Helpful features

When asked about the software features that they looked for or appreciated, experts referred to their ability to fit within their workflows*: “[…] one of the nice things about [the software I choose] is it then also leads into another piece of software that when you are getting the very high-resolution structures, it really helps at the end of that process.”* (E03) Another expert identified interactive features as helpful to compensate for the lack of their programming experience: *“I think having myself been someone with not a lot of programming experience, […] what I like to use things to be interactive because that means that they [software engineers] have actually put thought into how people will use it as a tool rather than just as some software they have installed.”* (E05) Transparency and feedback updates are also very important. Experts found using the software without receiving feedback about their actions and system responses to their actions frustrating*: “[…] I really like to be able to specifically see what I am doing […] having everything explained well enough*

within the software is that, yes, you can follow tutorials, but also the software is confirming what you are doing along the way […] half the time I am using this software. I have no idea about what the thing is that I am actually doing. I am usually just following like a protocol of what someone else has done and hoping for the best, whereas I need the software to feedback to me. If something is not working correctly or why it is not working correctly, then I can try and figure out how to fix that problem." (E06) Another expert identified real-time feedback provided in an easy-to-understand and intuitive manner as helpful, especially when using opaque systems*: "[…] it is very intuitive. It uses a traffic light-like system. It is like if it is running, it will go green. If it is not, [it will] go yellow, and it will tell you why. So it is very important because you do not see what is going on under it."* (E11)

4.4.4.5 Practitioner's view: They will find it if they need it

The AI team lead said that there was a specific feature called a slider that could show the uncertainty score of the output. They also admitted that even though they liked the tool, the users did not like it that much: *"[…] there is a tool that I like using, but not everybody does like it, which is this sort of slider, where you can say, tell me the things you are 90% certain of, and it will highlight just the things that it is 90% certain of."* (P02). The AI leader was confused when asked how experts learned about this tool and how they could have used it. When asked if the information about this tool was included in the training booklet, they said: *"I can't remember."* (P02) They then looked through the booklet to find more information: *"It is a really good point, because we do. We do in this section. Yeah, actually, we do not. We do not."* (P02) Another practitioner, when asked about the explainability feature, said that it existed, but users needed to know about it to be able to find it: *"It is accessible. They have to know what they are looking for. Um, and many people do not […] they are able to evaluate the output without [it]."* (P09) Then the practitioner tried to show it to me but could not find it or make it work: *"You can watch as I do a few things and then we can get to the point […] actually the confidence that we want is not running today for some reason."* (P09) When asked, experts said that they did not know about this feature. They were not informed about it, but they were also preoccupied with trying to get the software working rather than searching for helpful features. The fact that practitioners did not think about communicating with experts about the

available explainability feature suggests that explainability was considered the experts' responsibility to discover it and know how to interpret it.

4.5 Discussion

This study and its analysis present an actionable strategy for tackling AI adoption issues among domain experts. The analysis has revealed practical barriers to AI adoption at the human level. These transcend the influence of software capabilities (e.g., confidence, explainability). Many of these barriers stem from inaccurate assumptions of experts' needs and attitudes. I argue that without eliminating them, experts cannot explore the software and benefit from the features, such as explainability. Currently, there is a lack of research trying to understand the aspects unrelated to the AI technical capabilities that stifle the adoption. The study reported in this chapter details the importance of users' learning hurdles when new systems are embedded in their workflows. Guiding users through practicalities (e.g., software installation), providing active support at the beginning, and establishing ongoing collaborative efforts are fundamental to increasing AI uptake responsibly. This means that experts would make greater use of AI systems and be equipped to use them more meaningfully, with deeper understanding and motivation. These observations also align with both TAM (Venkatesh & Bala, 2008; Venkatesh & Davis, 2000) and UTAUT (Kulviwat et al., 2007; Venkatesh et al., 2012) models, arguing that the factors influencing technology acceptance are driven by the balance between the perceived benefits and effort required. The study showed that this balance can be achieved by clearly communicating the system's capabilities and reducing required effort by providing initial and ongoing collaborative support. The following subsections expand on the observed barriers and recommendations to overcome them.

4.5.1 Barriers to effective AI planning

One of the issues leading to further barriers in software implementation and adoption is a misalignment of what understanding of the underlying AI/ML mechanisms experts need to have to use the software successfully. Experts admitted that they did not fully understand them, but with the proper support, they could learn and did not see it as a critical challenge. Not having to understand how AI/ML worked to be able to run the software was seen as one of the benefits of it.

On the other hand, practitioners saw basic AI/ML knowledge as a prerequisite for using the software responsibly and effectively. The difficulty in accessing the software and more complex training were seen as a threshold for using the system. The argument was that if they cannot use it, they should not, as they would not understand why they got certain results. However, the issues preventing experts from using these systems were not always related to their deeper understanding of AI/ML mechanisms but to the lack of basic programming skills necessary for software installation. That does not indicate that users do not have a minimal understanding of AI/ML's underlying mechanism or that they would be unable to evaluate the system outputs with sufficient support. This misperception affected the practitioners' outlook on support and training. Although training was often the first opportunity for users to try the system, learn about it, and get support, practitioners did not see it as an important part of the AI adoption pipeline. Practitioners planned the training sessions based on what they saw as essential information that experts needed. The post-training surveys were not analysed, and experts were not contacted for feedback or input. These training sessions were a formality and a missed opportunity to help users learn to use the software and understand its capabilities and limitations. In some cases, the training was presented in a manner that was difficult for experts to relate to. For example, it used computer science jargon and did not show how the software could be useful to experts' projects.

4.5.1.1 Recommendation: Develop collaborative training solutions

Organise the collaborative training based on two-way feedback. The training should be more interactive than simple practitioner-to-user training, and experts should actively participate by informing practitioners about their specific needs. Interviewed experts said that they did not find the existing training useful. None of them were able to transfer their learned skills into a real-life context. Thus, the training material should be planned by collaborating with several experts, making sure that it is understandable to them and relevant to their tasks. For training to be practical and engaging, the training situations should be realistic, i.e., consistent with the experiences the trainee would face in an actual situation (Magerko et al., 2005). The effectiveness of training can also be increased by tailoring it to the particular needs of an individual trainee (Magerko et al., 2005).

The interviews revealed that there was no follow-up or meaningful feedback after the training. To address this, the training should have two parts, one based on teaching and discussing the material and the other discussing the issues that arose when trying to use the software independently. Talking to experts about the issues would also show what training areas should be improved. Experts should be asked to solve a specific task, independently explore the task domain, and discover task solutions on their own before the second training session. This method is called active and exploratory training (Keith et al., 2010). It has been shown to be effective even when trainees lack motivation and ability (Keith et al., 2010). The idea behind this type of training is that it incorporates the concept that learners should be trained on meaningful and realistic tasks. This method also supports error management, including recognising and recovering errors (Frese et al., 1991). Individual problem-solving is expected to cause users to make errors, which would be a valuable source of feedback and could help them prepare to address them in the future (Keith et al., 2010). Interviews showed that practitioners tried to teach users some basics of AI/ML mechanisms to ensure they could safely use their developed tools, but this method had not been effective. Error management training could be used instead of covering the basics of AI/ML. Practitioners should also clearly communicate the limitations and risks of the software during the training.

4.5.2 Barriers to AI adoptions during the introductory period

Experts struggled to bypass the initial steps of installing and running the software. They were more likely to reject the system if they experienced issues during this initial stage, especially if the software did not have comprehensible documentation and if they did not receive initial support. Experts complained that practitioners expected them to be programmers and to *just know*, for example, what command line instructions to use. However, they did not seek help because, as scientists, they were embarrassed to struggle with basic tasks. Practitioners, indeed, did not consider that experts might not have these basic skills. They saw tasks, such as software installation, as a matter of fact. When introducing their software to potential users, they used one-off training workshops, predominantly focused on explaining higher-level details, such as underlying mechanisms or the data processing pipeline. Although documentation and contact information were usually available, practitioners did not seek experts' feedback on whether the available training or documentation

was helpful. There was also no active effort to offer support or follow up on experts' progress during the initial stage.

4.5.2.1 Recommendation: Offer active support during the initial stages

Provide support during the initial stage of AI adoption. Before the new technology is fully introduced to the experts' workflow, they should be given time to adjust to the new system. During this stage, experts should receive active support from assigned team members, who would guide them and help them install and run the new system. This should also establish contact between an expert and a person/team to whom they could communicate their questions. Experts should be guided to the updated and comprehensible documentation, which would be established in collaboration with them. Providing support when implementing new technologies could positively influence users' attitudes towards technology and their self-efficacy (Laganá et al., 2011).

Experts should actively seek practitioners' feedback and clearly communicate theirs during this stage. During this stage, communication between experts and practitioners could improve how technology fits within experts' workflow (Orlikowski, 2000). It could also reduce the stress of adjusting to the changes. For example, Park and colleagues (2021) showed that discussing practical solutions to their AI problems with the management team provided emotional support, lowered social burdens, and demonstrated that managers cared for them.

Experts should also be allowed to practice using the new system. Practice sessions could enable a learning-before-doing approach, enhancing later performance (Pisano, 1996) and increasing motivation to use and experiment with the new technology independently (Edmondson et al., 2001). Allowing users to test and reflect on the new technology has also increased its acceptance. A study exploring the learning difficulties of novice technology users revealed that they mostly valued the opportunities to experiment and explore the system freely and safely (Bohanec et al., 2017). An ethnographic study of technology implementation in different cardiac surgery departments reported that sites that implemented trials for technology use followed by reflection sessions improved the chances of successful implementation and formed a learning cycle (Edmondson et al., 2001).

4.5.3 Barriers to investing time to learn and explore AI systems

Practitioners argued that users were unwilling to participate in software design and development stages and just wanted a final result or expected AI to *just do its magic.* This felt unfair to practitioners: "*[…] we are expected to learn biology, but they [experts] are never expected to learn the computing side of things*". Interviewed experts were willing to invest time and effort in learning and improving the software if they knew that the system would be helpful for their project. Experts prioritised systems that showed evidence of being beneficial for their specific project. For example, they searched for videos on Twitter, where scientists would demonstrate software and explain its features using language and context that were relatable to them. Otherwise, they felt that they risked wasting their time. However, the capabilities of the systems were not communicated effectively to them. Practitioners were presenting the technical aspects of the software without framing its capabilities in a way that would be relevant to experts. They also assumed that experts would discover software benefits independently and then communicate them to their peers.

4.5.3.1 Recommendation: Clearly communicate the capabilities of a new system

When introducing a new AI system, clearly communicate how experts could benefit from it. Communicate not only what the system can do but what it can do for them as experts and for their projects. Use domain-specific language and avoid technical jargon. The interviewed experts were willing to invest their time and effort if they could see how they could benefit from using the system. This aligns with research showing that the relevance of the learning material and tasks is an important motivator to seek more knowledge (Jones & Petre, 1994).

Showcase the capabilities of the system by using practical demonstrations. One of the most convincing factors for adopting a new system was seeing other experts use it on a similar dataset or a project. For example, one of the experts said that they would invest time in trying the software and learning about it if they saw another scientist demonstrating it on Twitter. Thus, it could be beneficial to include practical demonstrations (not for teaching purposes) tailored to the experts during training workshops. Alternatively, short videos could be prepared in collaboration with domain experts and made available online or internally to show how the systems could be used to benefit them.

Begin talking about the software's potential capabilities with experts before it is fully developed. This can help determine whether experts' and practitioners' expectations align, better understand experts' needs, and avoid developing unusable systems. It could also help experts feel more in control and less averse to the technology (Dietvorst et al., 2015; Sellwood et al., 2018).

4.5.4 Barriers to AI adoption due to effective collaboration between experts and practitioners

The study revealed the importance of ongoing collaborative efforts for responsible AI adoption. Infrequent communication without a pre-defined plan and set meetings did not lead to established collaboration, even when teams worked in the same building. Without it, members of both groups were increasingly frustrated because of the expectations projected onto them. The results of the study showed that consistent and planned collaboration could resolve these frustrations. When experts and practitioners had to work together, they established mutual understanding, aligned their expectations, and learned how to ask questions and explain information that would be understandable and meaningful to the other party. After collaboration, experts were motivated to learn and troubleshoot independently and invest time in taking steps to improve the software. When two teams worked together because of the nature of the projects, experts actively got involved, and even when the software did not perform as they expected immediately, they knew that it was still in the stage of development. Instead of rejecting it, they gave experts feedback that could help improve the system. Expert engagement can help to define expectations for system performance and to identify contextual factors that are likely to change when the system is implemented in practice (Dorton et al., 2022).

4.5.4.1 Recommendation: Plan the collaboration

Implement the plan for collaboration between the experts and practitioners (or other stakeholders). For example, start by agreeing on the type (online, in-person), frequency, and minimum timeframe for collaboration. Collaboration does not have to be extensive in the time spent working together, but it should be planned and continuous. Having a collaboration plan could reduce the burden of having to initiate meetings. It could also change the dynamic of experts only contacting practitioners to report problems or when they already feel frustrated. It could also help to persist in

working together, even when collaborative efforts are not immediately beneficial. Our results showed that establishing an understanding between multidisciplinary teams required persevering through moments when the collaboration felt ineffective. Our results also revealed that having a collaboration plan improved engagement from both teams. On the other hand, poorly planned practices have been shown to complicate effective collaboration (Seeber et al., 2020).

Use continuous collaboration to improve feedback. Research shows that improving interdisciplinary communication skills is essential for effectively providing and receiving feedback (Hardavella et al., 2017; Zhou et al., 2021). Effective user feedback is necessary for generating, refining, and fine-tuning design options and alternatives (Hardavella et al., 2017). Feedback is also critical to framing arguments and balancing multiple perspectives and considerations in the design process (Hardavella et al., 2017). Collaboration practices, such as team discussions, should also be applied during the initial stages of system use with the goal of learning. Learning through discussion is an effective way of acquiring new information. Collaborating has been shown to create a dynamic learning system with humans in the loop (Syed et al., 2020).

Define collaborative goals for different stages of the process. The ability to work effectively in teams is driven by having a common goal and assuming shared responsibility for completing tasks (Mishra & Kereluik, 2011). For example, during the planning or design stages, collaborate on aligning expectations and trying to understand experts and their workflows better. This study showed that starting collaborative efforts before the software was developed helped meet users' needs better. Early efforts also made users more confident, independent, and persistent when testing the new system, troubleshooting it, and communicating issues and feedback. For example, one of the interviewed experts involved in a collaboration volunteered to test an unfinished version of a new software. Collaboration could also be applied in the mid-stages of software development, with a goal to demonstrate and discuss a preliminary version of the project. Involving users through collaboration is necessary for effective AI development (Inkpen et al., 2019). The argument that experts' needs should guide the development of the technology is in line with user-centred HCI research goals (Hartikainen et al., 2022). Including users in the design process is also the basis of the participatory design methods (Zytko et

al., 2022), which have been effective in various domains (Badillo-Urquiola et al., 2019; Saxena et al., 2020).

4.5.5 Explainability and trust

Interviews with experts showed that being unable to overcome initial barriers to using the software prevented them from even thinking about the aspect of explainability and trust. This suggests the importance of supporting experts using the software, especially in the training and initial implementation stages. Only then could the conversation about explainability be effective. One interviewed expert knew about the explainability feature, showing the accuracy metrics. However, they were keen to continue manually checking the outputs, as the explainability feature only showed the number, and the expert lacked the context to inspect it using their expertise. This aligns with the research showing the importance of context and relevance in explanations (Gu et al., 2020). Another expert expressed concern that explainability would become another complex feature of the software they must learn to use. Brennen (2020) reported similar experts' views on explainability. This also suggests that the first impression of software not being accessible and designed for experts' needs and abilities could negatively influence their trust and perceptions of explainability. These findings emphasise the importance of setting the foundations for effective explainability.

4.6 Further Implications

The implications of these findings extend beyond the recommendations for increasing expert acceptance of AI tools. This study revealed that the growing pressure and necessity to use AI systems can put a significant emotional burden on experts. Introducing new technologies into experts' workflows without appropriate support or preparation negatively affected their emotional well-being and reduced the time they could spend on their projects. The interviewed scientists were experts in their area of research who had access to the best research equipment, creating the potential for making ground-breaking scientific discoveries in life sciences. However, they were spending a lot of time figuring out how to use software and were unable to process the available data. Experts were frustrated and stressed during the interview, a few even acknowledged that this process negatively affected their

mental health and confidence. Addressing the recognised issues could reduce this stress and allow experts to focus on their expert tasks and progress with their work.

Moreover, AI developers would also benefit in the long term. Applying the proposed suggestions could help calibrate experts' expectations, motivate them to learn about the AI system, advocate it to their peers, communicate their feedback and be more independent users. The interviewed practitioners referred to the lack of resources, such as time, that prevented them from providing more support for experts. However, they admitted that fixating on the technology and presenting it to experts only when it was developed was ineffective. This led to cases where their software was unused and had to be abandoned. Our study suggests that finding initial support time by reprioritising tasks could pay off later, as practitioners would have to spend less time troubleshooting and advocating their products.

This study also emphasised the lack of appreciation for experts' limited computational knowledge and skills. The study showed that experts struggled with practical aspects, such as command lines and installation processes. However, they need these tools, as processing their data manually is no longer an option. This study showed preparation and ongoing support need to be prioritised when introducing new tools. Otherwise, there is a risk that the application of expert skills will be limited across domains that now require the adoption of AI. These findings could facilitate research approaches, such as participatory design, requiring meaningful user engagement and collaborative efforts between stakeholders. The proposed steps could motivate users to get involved and invest their time. Following these steps towards better collaboration could help users and practitioners align their expectations, build communication skills, and understand how to communicate effectively.

4.6.1 Limitations and future directions

This study focused on a specific domain and two teams working within the same organisation and environment. A future study involving more participants of varying levels of task and domain expertise is necessary to explore further barriers to AI adoption arising at various stages. These findings provide meaningful insights regarding AI adoption among experts that could be generalised across different areas and domains. However, further studies are needed to test the proposed

recommendations in domains of varying risk, time pressure, and the extent to which human experts rely on AI.

This study recognised an assumption that domain experts will have a certain level of practical computational skills. The lack of appreciation for experts' limited computation knowledge prevented them from adopting AI tools effectively. Simply using minimal technical knowledge as an entry requirement or a safety barrier was not a viable option. Experts often do not have a choice but to use AI to exercise their expertise. The pressure to use new technologies will only increase with further advances in AI and data collection methods. Researchers should think of ways to support and motivate experts to learn about AI systems. However, they should also explore ways to inform developers on how to communicate information about their developed systems effectively and support users more strategically. Future studies should examine ways to support mutual learning processes through collaboration.

4.7 Chapter overview

Explainability solutions are not useful unless users adopt AI technologies effectively and use them to support their work. Explainability research often considers explainability effectiveness and usability, assuming that experts can use the AI system. The practical hurdles users face when AI systems are embedded in their workflows are understudied. This study revealed that guiding users through mundane practicalities (e.g., software installation), providing active support, and establishing ongoing collaborative efforts are fundamental to increasing AI uptake. Easing experts' journeys of using new AI systems and establishing collaborative training and support practices could, in turn, increase their motivation to engage with explainability practices. Explainability could become a part of the conversation and be tailored in collaboration with experts. The actionable steps presented in this chapter aim to remove the key barriers separating experts from the systems they need to use. Here, I propose strategies that could address the issue of AI adoption among experts and pave their path towards the effective use of AI, where explainability could be a useful feature. These strategies are: i) developing collaborative training solutions; ii) actively supporting experts during the initial stages of software use; iii) communicating the software's capabilities in a relevant to experts' language; and iv) following predefined collaboration rules. Following these recommendations could make experts more accepting of AI systems and increase

their motivation to understand the technology they use, advocate it to their peers, and be more independent and confident users. Moreover, it could help them to be more specific in communicating their feedback to practitioners. This study also showed that experts who struggled using AI tools and were frustrated about installation, guidelines, and inability to use them did not even think about explainability. Adequate support and strategic team collaboration could make experts more conscious and responsible AI users and allow them to benefit more from new technologies. These findings could facilitate research approaches, such as participatory design, which requires meaningful user engagement and stakeholder collaboration to improve AI technology's contextual fit and usability. Overall, this chapter presents steps for building a foundation for explainability through supportive, collaborative implementations of AI technologies.

Chapter 5 Tailored explainability for domain experts: Strategies to enable expertise in AI-driven decision-making

Chapter 5 explores how XAI could be tailored to enable expertise in AI-driven decision-making. It explores i) what strategies experts use to make decisions and reason in naturalistic environments, ii) how the introduction of AI could disrupt these strategies, and iii) how explainability interfaces could be designed to align with expert cognitive and contextual needs. This chapter is based on the review of HFE research literature on expert decision-making in dynamic contexts. To my knowledge, this is the first attempt to extrapolate from the HFE literature to inform the XAI research. The HFE discipline was chosen as its research on human-automation interaction has many parallels with human-AI interactions in expert contexts. Based on the literature review, I developed a conceptual Expertise, Risk and Time (ERT) framework intended to inform explainability interface design choices depending on the three dynamics. This framework was then used in a simulated journalism case study to illustrate its real-world applicability. Finally, the ideation workshops were organised to brainstorm a list of concrete design solutions fitting the framework.

This chapter explores the aspect of *enabling expertise* – the first of the three properties of the holistic XAI approach proposed in this thesis.

5.1 Chapter introduction

The research literature discussed in the previous chapters showed that experts often cannot stay in the loop effectively, especially when the technologies they use are overly complex or opaque (Young et al., 2019; Yu et al., 2017). This leads to either aversion to AI (Dietvorst et al., 2015), overreliance (Gaube et al., 2021), loss of agency (Cranefield et al., 2023), high mental workload (Lindgren, 2023; Wickens et al., 2013), and inability to use AI systems effectively (Amann et al., 2022; Gaube et al., 2021; Micocci et al., 2021). These issues persist or are amplified even when AI outputs are supported by explanations (Bansal et al., 2021; Kaur et al., 2020). One of the downsides of available explainability approaches identified in the previous chapters is the lack of support for expert decision-makers. Although they are one of the critical stakeholders for XAI, available techniques are not designed with the intention of preserving and enhancing their expert knowledge and skills. A few studies involved domain experts in explainability experiments, trying to understand

how they respond to different explanations (Bussone et al., 2015) and to understand potential cognitive biases occurring in AI decision-making contexts (Wang & Yin, 2021). However, there is a knowledge gap in understanding how explainability could support expert decision-making and preserve the expertise developed through years of experience performing specific tasks, making decisions, and analysing information.

The XAI research direction has recently shifted to more user-centred approaches and understanding of how people use explanations (de Graaf & Malle, 2017; Eiband et al., 2018) and what could be learned from disciplines such as psychology, philosophy, and cognitive sciences (Beaudouin et al., 2020; Hoffman et al., 2018; Miller, 2019). However, there is still no clear strategy for achieving explainability in expert domains. There are also no design guidelines that would advise which explainability interface design approach would be the most suitable in which situation depending on the decision makers' needs and contextual factors. Exceptions include a set of usability guidelines by Amershi and colleagues (2019) and XAI Question Bank by Liao and colleagues (2020). However, the former, although relevant, is not specific to explainability and the latter is based on interviews with UX practitioners and designers rather than domain experts. There is a need for guidelines that would demonstrate how explainability could be used to support domain experts, for example, what to explain and how to display explanations in the interface, as well as how to account for real-world constraints.

A better understanding of experts' needs and preferred ways to receive and analyse information when making decisions could inform the design of explainability interfaces and explanations. Beyond the techniques and types of explanations outlined in Chapter 2, explanations can also vary in aspects, such as complexity (Lage et al., 2019), completeness and the amount of detail (Bussone et al., 2015), interactivity (Madumal et al., 2019; Weld & Bansal, 2018), and availability of feedback (Smith-Renner et al., 2020). Tailoring these and similar aspects of explanations to the user or a context might lead to increased usability and effectiveness of explanations (Schaffer et al., 2019) and reduced cognitive load (Naiseh et al., 2021). However, effective XAI tailoring requires establishing an understanding of what aspects should shape the personalisation of explanations and how (Ras et al., 2018). Even when proposing tailored explainability approaches, researchers tend to base them on aspects such as domain characteristics (Gilpin et

al., 2018), relevant legal requirements (Beaudouin et al., 2020) or task goals (Hind et al., 2019) rather than naturalistic ways of expert decisioning. This study attempts to define higher-level decision-making and sensemaking strategies that are domain-agnostic and could enable experts in AI-supported decision-making.

5.1.1 Publications

This chapter is based on the Journal of Responsible Technology publications *Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable* (Simkute et al., 2021) and *Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context* (Simkute et al., 2020). I researched and wrote both papers, with the exception of the journalism case study in the *Explainability for Experts: A Design Framework for Making Algorithms Supporting Expert Decisions More Explainable* (Simkute et al., 2021). This part was done in collaboration with an experienced journalist, Dr Bronwyn Jones. The chapter is finalised with a list of design suggestions to support the ERT framework guidelines. These suggestions were refined from the results of the design workshops with interface designers and postgraduate design students.

5.2 Methodology

Following the second stage of the literature review, it became apparent that tailoring explainability to support domain experts requires understanding them. In particular, experts must be understood in terms of how they make decisions and what information can support their ability to form meaningful trust and stay in the decision-making loop. During the third stage of the review, I aimed to explore expert decision-making strategies and how the introduction of AI and various contextual factors influence them and answer the following research questions:

1. What decision-making and sensemaking strategies are unique to experts?
2. What are the critical contextual factors influencing expert decision-making, and how does decision-making change under those factors?
3. Extrapolating from the human-automation interaction studies (HFE domain), how could the introduction of AI stifle the naturalistic decision-making strategies of domain experts?
4. How can explainability interfaces be designed to support experts' naturalistic decision-making strategies?

I also aimed to develop a conceptual framework that would help to tailor explainability interfaces to support experts' information needs under various contextual circumstances and accommodate workflow disruptions due to the introduction of AI.

The HFE discipline was chosen for this review stage as it has studied experts and their interactions with automation over three decades. First, an initial scoping review was conducted, exploring key research papers within HFE that were the most influential in the research area of expertise and decision-making. "Expertise" in combination with either "sensemaking", "reasoning", "mental models", or "decision-making" were selected as search keywords and were used to find further publications. I focused on the most cited publications, which I further filtered using set exclusion criteria. I excluded papers that did not investigate human aspects in the decision-making context or research that was not transferable in the AI-supported decision-making context. The publications that focused predominantly on practical aspects or specific interfaces rather than the interaction with the technology were also excluded.

5.3 Expert decision-making: Human Factors Engineering perspective

The HFE research literature is rich in empirical knowledge and theories on how experts make sense of available information and how they proceed with their decisions in their work environments (Klein, 2003). This research provides insight into how expertise shapes expert decision-making strategies, responses to workflow changes and various contextual factors, such as risk and time pressure. The high focus on the Cognitive Psychology perspective in HFE research also provides an understanding of expertise-related biases (Stuart et al., 2007). Specifically, extrapolating from the HFE research can help to recognise what heuristics are triggered by various circumstances and how they influence expert decision-making. The knowledge of expert behaviours, contextual factors, and cognitive biases occurring in specific contexts could guide explainability design and increase the applicability of existing XAI methods.

Moreover, the HFE research field has been actively studying a factor of automation. This research explored the impact of automation when it was first introduced to areas such as air traffic control, including unanticipated side effects of automation (e.g., reduced productivity, difficulties supervising automation, and

automation bias) (Bainbridge, 1983; Parasuraman & Riley, 1997; Warm et al., 2008). It also focused on the development of interfaces tailored to specific tasks to support, enhance, and preserve expert skills (e.g., Ecological Interface Design) (Rasmussen & Vicente, 1989). The HFE research findings give a unique opportunity to inform the relatively young research area of expert-AI interactions that have many parallels to the expert-automation interactions. By extrapolating from the HFE research literature, we can better understand the challenges of introducing AI in expert domains and potential adaptation strategies. The knowledge of how introducing new technologies might disrupt expert decision-making can help plan how explainability could counter the challenges caused by these disruptions.

The literature review revealed three main dynamics significantly influencing decision-making and sensemaking strategies. The first one is the level of expertise, which influences the specific decision-making strategies that distinguish them from novices. For example, experts can make intuitive decisions and recognise patterns and irregularities in data (Klein, 2003). They prefer exploring available information freely and flexibly (Klein et al., 2010) but are susceptible to overconfidence and similar biases (Eva & Cunnington, 2006; Kahneman & Klein, 2009). The second factor is a contextual risk, which increases cognitive workload and the risk of heuristic-driven decisioning (Orasanu, 2010). Experts can apply specific strategies in high-risk contexts to enable analytical deliberation. For example, they can use rule-based protocols, schemas, or additional contextual information, which can help challenge their intuition (Sibbald et al., 2013). The third factor is time pressure, which further increases the cognitive workload and challenges analytical decision-making. When pressured by time, experts need to recognise patterns and irregularities quickly (Klein et al., 2010), but if they have too many information resources available, they can fall for overutilisation bias (Gonzalez, 2005).

5.3.1 Expert decision-making

According to the dual processing theory, people use two parallel systems when making decisions: fast, intuitive, subconscious reasoning, with little use of working memory (System 1), and analytic, deliberate, and recursive process of reasoning (System 2) (Kahneman, 2006; Kahneman & Klein, 2009). One of the critical aspects differentiating novice and expert decision-making is the ability to respond to decision-making situations intuitively using System 1 effectively. Experts can make decisions

without evaluating every available option or potential outcome through deliberate practice and extended exposure to the specific task (Hutton & Klein, 1999). By using intuitive thinking, experts can effectively reduce cognitive load when making decisions compared to novices. HFE research has shown that experienced decision-makers follow intuition with little conscious deliberation and rarely consider multiple options (Hutton & Klein, 1999). When experts have to test a hypothesis analytically using System 2, they use schema, a set of domain-specific knowledge they possess. For example, expert doctors have been shown to use knowledge templates built on previous experiences when making diagnostic decisions (Sibbald et al., 2013). Instead of considering all possible variables, they intuitively recognise the key ones (stored in their knowledge templates) that are important to the diagnosis (Ross et al., 2006). Novices, on the other hand, use less efficient strategies, such as testing one hypothesis at a time (Coderre et al., 2003).

Intuitive decisioning does not mean that experts make important decisions carelessly. Instead, expert intuition can be an essential asset in complex decision-making situations. HFE research suggests that experienced decision-makers often better recognise unusual cases when they trust their intuitions than when they engage in detailed analysis (Klein, 2015a). For example, Benner and colleagues' (1992) showed that experienced nurses who assessed the situation using only their expertise was more effective in identifying unusual cases and critical illness prevention information that otherwise might have been ignored. On the other hand, novice nurses, who relied on a protocol-based checklist, could only diagnose but not anticipate and prevent illness. Expert nurses could also find solutions faster by generating fewer options that needed consideration, while novices had to produce several options and conduct analytical comparisons of them (Benner et al., 1992).

5.3.1.1 Experts are vulnerable to overconfidence and illusion of validity

However, a high level of expertise can lead to the "illusion of validity", an unjustified sense of confidence and, hence, failure to evaluate different possibilities (Kahneman & Klein, 2009). Experts' failure to challenge their intuition has been shown to lead to missed automation errors (Eva & Cunnington, 2006). Introducing AI could increase the risk of overreliance on intuition as it removes essential contextual information and only provides an output. When provided information (e.g., automation aid suggestion) is stripped of its usual contextual information (e.g., smells and images),

it has been shown to bias experts towards the most salient option (Kahneman & Klein, 2009) or rejection of the output if it is not in line with their intuition or preferred way of working (Westin et al., 2016). When experts receive only an output, they do not have much information to inform their intuition and are left with limited means to challenge it (Klein et al., 2006a). The introduction of automation has been shown to disrupt experts' ability to apply their natural decision-making strategies, especially if they were only shown the final output of already processed data (Klein et al., 2006b).

5.3.1.2 Experts' intuition is driven by pattern recognition in the contextual information

According to the recognition-primed decision (RPD) model, experts' intuition is driven by their ability to create mental patterns in data (Klein, 2003). When shown the information, they notice cues and intuitively link them to other cues they expect to appear next based on their experiences (Klein et al., 2010). Experts are particularly sensitive to the context and thus are better at noticing features of situations that could have potential implications (Klein et al., 2010). They quickly and subconsciously recognise patterns and cues in situations (or data sets) and intuitively link them to other cues they expect to appear next (Schmitt & Klein, 1999). According to this model, high-expertise decision-makers rely on past experiences to recognise cue patterns that allow them to understand and evaluate the problem or information. Cue recognition triggers retrieving a response drawn from a similar experience with matching cue patterns (Orasanu, 2010). Seeing the noisy data, not only the main trends, guides the experts' intuition and triggers pattern recognition, allowing them to know which cues to monitor and which are essential or doubtful (Ross et al., 2006). Experts under time pressure are likely to engage in RPD decision-making (Klein et al., 2010). However, RPD decisions are unlikely when a decision requires justification, as intuitive decisioning is hard to articulate (Orasanu, 2010).

Automation has been shown to interfere with experts' intuitive pattern recognition. For example, when experts could not access supporting data, they could not use their expert ability to notice and test inconsistencies (Klein, 1993). In another study, expert weather forecasters refused to use algorithmic metrics provided by a computer system, as these were "too smooth" and only revealed the main data trends but not the *noise* (Stuart et al., 2007). The expanded data set can guide experts' intuition and trigger pattern recognition. It can help them recognise which

cues are essential or doubtful and should be monitored. An explainability interface should uncover cues and patterns that can lead to deeper analysis and decision evaluation. Moreover, expert users might be reluctant to adopt intelligent systems if they cannot use their intuition and understand the system's limitations and nuances, even when the suggestions made by algorithms are in line with their predictions (Hilburn et al., 2014).

5.3.1.3 Expert information-seeking and search strategies are not neat and orderly

Experts seek additional data in a free and explorative way, especially when the situation is novel, or they need to contest and investigate the provided output (Schmitt & Klein, 1999). Their search strategy for relevant information that could help to reduce uncertainty depends on the material they encounter during the search. Experts decide what further information they need during decision-making, not in advance. Thus, their search is dynamic and unpredictable rather than neat or orderly (Klein, 2015a). Experts do not like to be constrained by strict search categories and ways information is provided. When experts cannot freely explore data or only receive outputs without relevant supporting information, their analysis suffers (Klein et al., 2006b). Participating in active search, adaptation, and mental model-building processes allows experts to effectively exercise their expertise and maintain motivation due to the increased agency (Klein et al., 2006b). Not allowing experts to explore additional information freely but providing them with too much data at once can reduce accuracy and increase overconfidence in their decision-making (Klein et al., 2007). To achieve the best results, experts should be allowed to form an early conclusion based on their intuition, then deliberately test the initial conclusion and seek more information using their preferred methods (Klein et al., 2007).

5.3.1.4 Experts struggle to apply their expertise when their workflow is disrupted

Introducing intelligent systems can disrupt experts' ability to use their expertise effectively (Klein et al., 2006b). For example, introducing automation stifled experts' ability to effectively use their skills due to disrupting their naturally occurring sensemaking strategies (Klein et al., 2006a). According to the Data/Frame theory, the natural sensemaking process begins with an initial viewpoint of the topic called a frame. Sensemaking consists of questioning that initial frame using data, then challenging the data-based explanations, elaborating on the frame and adding details to it (Klein et al., 2007). The Data/Frame theory proposes that the most

effective way to make decisions is to recognise the frame early in the process and commit to it while testing it. This process is disrupted when information is provided as an output, and an individual cannot perform these steps (Klein et al., 2007). Klein et al. (2006a) warned that “spoon-feeding” experts with data interpretations can be frustrating and demotivating. It could also be counterproductive, as experts could not recognise, commit to, question and reframe the initial frame. Experts should be motivated to challenge the frame (i.e., an output) and to recognise errors in automation performance. Experts in real-life decision-making situations have been shown to actively challenge information when they recognise the potential that the output is inaccurate (Sterman, 1994). In AI-supported decision-making, the expert user should be provided with additional information (e.g., accuracy measures), which could work as a motivator to challenge the results.

Automation can disrupt experts' workflows, reducing effectiveness or disabling the application of expert decision-making strategies learned with experience (Sterman, 1994). Even when the task maintains the same logical structure as before the automation was introduced, contextual changes might prevent existing skills from being transferred to the new environment (Sterman, 1994). Subsequently, when introduced to that new environment, both novices and experts are likely to rely on their common sense or biased thinking, thus underestimating other aspects and only searching for consistency with their existing beliefs, leading to confirmation bias (Nickerson, 1998). The poor contextual fit of the automated systems also increases the risk of automation bias and algorithmic aversion (Elwyn et al., 2013). Research shows that even when automation is designed to support a specific task, it can fail if contextual factors are ignored (Nickerson, 1998).

5.3.2 Decision-making in a high-risk context

High-risk situations can be defined as situations in which decisions are likely to have significant consequences and/or can result in discrimination, loss of credibility, or even loss of life or property, as well as situations in which expert decision-makers face high performance and social pressures (Orasanu, 2010). In high-risk contexts, expert decision-making strategies are dependent on aspects that can accelerate stress and increase cognitive workload, in particular, the uncertainty of information (Orasanu, 2010), perceived lack of control (Breznitz et al., 2013) and time pressure (Perlow et al., 2002). In unfamiliar high-risk situations, where information is

ambiguous or incomplete, expert decision-makers are believed to search for cues that link the current situation to their past experiences (Orasanu, 2010). The influence of stress is particularly high when these information cues are unclear and cannot be recognised, assessed or matched (Orasanu, 2010). In these situations, expert decision-makers proceed with the cognitively demanding process of generating and matching multiple solutions while simultaneously considering potential consequences (Orasanu, 2010). More precisely, in high-risk contexts, they reduce uncertainty by matching the situation with similar past experiences, then generating and evaluating potential options serially (one at a time) and, if time allows, mentally stimulating potential scenarios (Lipshitz & Strauss, 1997). This strategy places a high cognitive load on the expert's working memory, as multiple goals and strategies must be held in it while simultaneously retrieving and evaluating various constraints (Orasanu, 2010).

5.3.2.1 Challenged analytical reasoning in high-risk contexts

Because decision-makers have to actively infer from available information, make predictions, and fill the gaps of missing information, they might be prone to overestimate their abilities to do so accurately. Expert decision-makers generally show overconfidence in their decision-making and forecasting abilities (Kahneman, 2006) and make errors by overestimating their impact on the outcome (Langer, 1975). Thus, they are susceptible to the illusion that their predictions are correct and over-commit to their choices (Einhorn & Hogarth, 1981). To avoid these biases, in high-risk situations, experts might have to employ thorough and analytical decision-making strategies to slow down their decision-making process instead of making fast and intuitive decisions (Svenson, 1979). However, to apply analytical thinking, they need support and structure that would allow them to weigh available options or decision scenarios and lower cognitive load related to high-risk (Svenson, 1979). Experts show better performance and ability to evaluate available information in high-risk situations when using a rule-based protocol or a checklist to reduce ambiguity (Orasanu, 2010). Research in a medical setting showed that the effects of uncertainty could be minimised by helping expert decision-makers match various situations to a rule-based protocol (Dobrow et al., 2006). This study showed that support tools such as decision principles and evidence hierarchies were essential in revealing important modifiers experts could recognise and use for decisioning

(Dobrow et al., 2006). Decisions that follow a set of rules or a specific checklist are less susceptible to stress, as they are made by linking the cues and patterns to examples or past instances and allow decision-makers to retrieve potential solutions from long-term memory (Orasanu, 2010). Providing rule-based protocols and checklists can lower cognitive load and enable analytical decision-making (Orasanu, 2010).

5.3.3 Decision-making under time pressure

Decision-making strategies are highly affected by the time available to process the information and develop the best solution. Time pressure has been shown to influence the pace and quality of decision-making (Perlow et al., 2002) and increase cognitive load (Oliva & Sterman, 2001). It further complicates decision-making when the cognitive load is already high, for example, in high-risk contexts or when the information is insufficient or incomplete, and experts struggle to recognise helpful cues and data patterns (Svenson, 1979). Experts under high time pressure adapt more straightforward decision-making rules and observe and assess information by dimensions (Payne et al., 1992). High time pressure does not allow a thorough and detailed analysis of potential outcomes, increasing the risk of automation bias or "illusion of validity". When decisions must be made quickly due to the limited time available or the high cost of delay, the decision-making strategy changes from analytic to intuitive. This way, decisions are made without exhaustively searching relevant information (Svenson, 1979). Expert decision-makers pressured by time are likelier to take "shortcuts" when using analytical thinking. They tend to process information serially by generating and evaluating one option at a time until the most reasonably fitting is accepted (Klein et al., 2010).

5.3.3.1 Too much information can lead to biased decision-making

When decision-making is constrained by time, the amount of information presented can influence its effectiveness by regulating the cognitive demands of the task. For example, providing multiple alternatives does not help to reach more valid or reasonable decisions. Due to sequential processing, evaluating alternatives can be cut short once an acceptable option is met, leaving the rest of the alternatives unconsidered (Hutton & Klein, 1999). When making decisions under time pressure, experts perform better when they receive less information and have fewer alternative

options (Payne et al., 1992). Receiving too much information or alternative resources can lead to ineffective decision-making (Wearing, 2004). Firstly, there is a limit to how many resources can be assessed effectively and how much information can be processed under time pressure. Thus, providing access and guiding the decision-maker to the vast amount of information can be counterproductive (Wearing, 2004). In time-pressured contexts, the decision-maker might feel an urgency to use all the available resources, such as information gathering, opportunity for action, or communication input (Wearing, 2004). This is mainly believed to be due to the overutilising bias – the tendency to exploit all resources outside conscious awareness (Reason et al., 1997). Decision-makers, including experts, are intended to believe that they can effectively manage the information and resources available. However, they do not appreciate the limitations of their ability to regulate the related cognitive workload (Wearing, 2004). The overutilisation of resources does not stop when the cognitive system of a decision-maker is overloaded, damaging their ability to make effective decisions. Thus, access to multiple information sources can be disadvantageous (Seagull et al., 2001).

Besides overutilising bias, the tendency to overuse available information comes from other general biases. For example, under time pressure, commission errors are preferred over omission errors, meaning that decision-makers are likelier to make mistakes when proceeding with action rather than due to delay or inaction (Kerr et al., 1996). Acting instead of waiting, even if ineffectively, also brings an illusion of control over the task, the sense of achieving some results and a sense of greater self-competence via activity (Schmitt & Klein, 1999). Due to the illusion of control (Duhaime & Schwenk, 1985), decision-makers make errors by overestimating their abilities and impact on the outcome (Langer, 1975). They may assume that through additional effort, they can make their strategy succeed should problems arise (Langer, 1975). Lastly, when presented with multiple resources, decision-makers express overconfidence bias and overestimate the amount of information and how fast they can effectively manage it in their working memory (Binmore et al., 1993).

5.3.4 The role of AI

Recent studies have demonstrated that expert decision-making is often disrupted when AI systems are introduced (De-Arteaga et al., 2020; Gu et al., 2020). This review helps to understand this disruption better. For example, the increased

cognitive load can explain poor expert performance when working with AI support. As with the introduction of automation, the introduction of AI systems might increase the unfamiliarity of the situation due to the workflow change. If the system is opaque, the uncertainty about the trustworthiness of the AI outputs can increase the cognitive workload, which is already high in these situations, further disrupting the analytical reasoning abilities and increasing the risk of biases and overlooked errors (Haldane & May, 2011; Parasuraman et al., 1993). AI-provided outputs also lack the context that could help experts effectively search for the cues that would allow them to reduce that uncertainty (Dobrow et al., 2006). Especially in high-risk situations, experts might be eager to seek the tools that would help them match the current situation to a rule-based protocol, checklist, or case example (Dobrow et al., 2006; Orasanu, 2010). However, AI systems often only provide a prediction or recommendation, disrupting experts' ability to test their intuition and notice important data patterns. Decision-makers using AI tools report feeling that the outputs they receive do not allow them to see the complete view of the situation. For example, medical experts resisted AI support because its static data outputs would not allow them to see the complete picture of their patients (Yang et al., 2019). They preferred seeing which factors were most influential towards the predictions, as this would allow them to see which factors were modifiable so they could plan future actions and interventions (Yang et al., 2019). Experts also express dissatisfaction if the AI system does not allow them to see the raw features of the data, which is necessary to interpret it the way they had been trained without algorithmic support (De-Arteaga et al., 2020). AI also limits experts' freedom to explore data, how they feel guided by their intuition, or how they are used to doing it. The lack of freedom to explore additional data has also been criticised by experts in recent studies of AI-supported decision-making (De-Arteaga et al., 2020; Yang et al., 2019). The HFE studies also showed that experts' ability to apply their sensemaking and decision-making strategies is linked to their familiar workflow (Elwyn et al., 2013). The introduction of AI disrupts it, leaving experts feeling restrained by the contextless nature of the predictions and the new structure of their task (Altmann et al., 2014). Failure to appreciate the context in which decisions are typically made without algorithmic support is one of the main reasons predictive systems fail in practice (Wagner, 2019). Poor contextual fit can make experts feel limited, which could result in them refusing to rely on the AI's predictions (Khairat et al., 2018; Yang et al., 2019). For

example, Veale et al. (2018) interviewed workers in public sector organisations and showed that whether experts interacted with AI systems and whether they relied on them depended on their fit within their natural workflow and organisational context. When unable to apply their decision-making strategies, experts tend to turn to their old methods (even if less effective) instead of relying on algorithmic predictions (Lee et al., 2017). The explainability interface should be carefully crafted, considering the particularities of the experts' decision-making processes. How information is presented can either help or further disrupt the expertise in decision-making (Klein et al., 2006b). Knowledge of how expertise, risk, and time pressure influence decision-making can help to inform explainability design.

5.4 Contextual ERT framework for explainability interface design

The review of HFE research literature concerning expertise, decision-making, and sensemaking strategies revealed the critical ways levels of expertise, contextual risk, and time constraints influence sensemaking strategies, cognitive biases and attentional resources in a decision-making context. The three dynamics that shaped the framework are (1) *the level of expertise* (i.e., prior experience in making decisions without any algorithmic support), (2) the *level of risk* (i.e., the cost of error) and (3) *time pressure* (i.e., the cost of delay, or given time to complete the task). Based on the reviewed literature, I mapped decision-making strategies that were likely to be employed under different combinations of these dynamics and developed the ERT explainability framework (Fig. 1). The Framework divides the decision-making space into four main sections, each representing a combination of high and low levels of expertise and risk. Each section is then moderated by the level of time pressure in each context, dividing the decision-making space into eight segments. Each segment represents different decision-making and sensemaking strategies, potential cognitive biases, and a risk of increased cognitive load. The framework is intended to help match these aspects with suitable design approaches and characteristics of explanations and explainability interface design.

The ERT explainability framework is suitable for deployment and iterative development. Its long-term goal is to support the development of effective design heuristics for explainable interface design in various contexts. By offering three clear dynamics, I introduce a framework for designers seeking to scope out the explainability requirements in a given context.

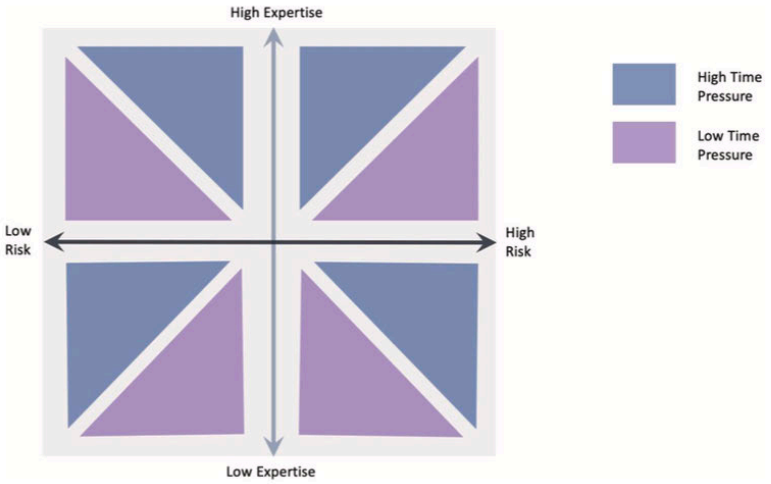


Figure 1. The ERT explainability framework divides the decision-making space into four main sections, each representing high and low levels of expertise and risk. Each level of time pressure then moderates each section.

5.4.1 Dynamic 1: Level of expertise

The ERT Framework distinguishes between design features for experts and novices. An expert in this context is a trained professional with experience in a specific domain whose expertise is a result of "a rich instrumental experience in the world and extensive and deliberate practice and feedback" (Hoffman et al., 1995). Whether a person is an expert can be established by their reported familiarity with the task (Schaffer et al., 2019) or the type and quantity of experience working in a specific domain (Hoffman et al., 1995). The ERT framework only separates two levels of expertise: high level, i.e., the decision-maker is an expert, and low level, i.e., the decision-maker is new to the task and domain. Although there are varying levels of knowledge in every domain, research shows that only after becoming an expert decision-maker can one begin to employ significantly different decision-making strategies than people in other knowledge categories (Dreyfus & Dreyfus, 1986). Progression to expertise happens when a person advances from a superficial and literal understanding of a problem to an articulated, conceptual, and principled comprehension (Hoffman et al., 1998). The ERT Framework does not consider the experience in using AI systems. It is designed assuming that domain experts are new to the technology and have a limited understanding of it.

Failure to consider decision-making strategies applied by experts when desiging system interfaces, could lead to compromising their ability to exercise skills and recognise particularities they learned to notice over years of accumulated

experience. Experts also resist to use AI systems if they cannot apply their intuition and understand the system's limitations and nuances, even when the suggestions are in line with their predictions (Hilburn et al., 2014). However, a system interface designed for an expert could overwhelm any novice decision-makers who use it, leading to errors and not allowing the development of expertise.

5.4.1.1 Uncover cues and patterns and enable mental simulation of features

Experts' analytical abilities are compromised when they cannot access supportive contextual information that would help them recognise cues and patterns and use their intuition (Billings, 1991; Lee & Seppelt, 2009; Moray et al., 1986). For example, expert weather forecasters refused to use algorithmic metrics provided by a computer system, as these were "too smooth" and only revealed the main trends of data (Stuart et al., 2007). The lack of contextual information has also been shown to increase the cognitive workload (Bainbridge, 1983). An explainability interface should support expert decision-making by providing relevant contextual information to help uncover cues and patterns, leading to analysis and decision evaluation. For example, via saliency maps showing a bigger picture and diagrams revealing data patterns, including noise and unusual interactions, that do not necessarily qualify as predictive. This information is necessary to trigger pattern recognition and help experts notice aspects that might indicate an error or trigger them to investigate the AI outputs further. After recognising the potential cues, experts can mentally simulate various scenarios, linking them to their previous experiences (Klein, 1993). Interactive interfaces could help experts perform these simulations when evaluating available information and determining the best next action (Cheng et al., 2019).

5.4.1.2 Support flexible information search

When an expert's ability to explore data is restricted or when they are unable to actively question and investigate the output, their analysis suffers (Klein et al., 2006b). Thus, a tailored explainability interface should support experts' need for flexibility in accessing data and enable their active involvement in information exploration. For example, it could allow them to interrogate an explanation or an output. An interface could also provide additional agency through feedback and modification features or by allowing experts to correct errors and gradually improve and customise the system (Kulesza et al., 2015). Experts' preference for flexible information search and exploration could also be fulfilled by providing interactive

search features (Weld & Bansal, 2018) and introducing refinement tools that could guide the search process (Cai et al., 2019).

5.4.2 Dynamic 2: Level of risk

Secondly, explainability design heuristics should be tailored depending on the level of risk assigned to the task. However, the complex nature of high-risk situations makes it difficult to draw predefined domain-agnostic risk criteria. Here, the high-risk context is considered as one that involves decisions that can have major consequences for the decision-maker or other involved and affected stakeholders or one that has a high cost of error. The level of risk might depend on aspects such as i) the extent of the consequences and the chain of people affected by the decision, ii) the temporality or permanency of the consequences (i.e., how long will these consequences be present and whether there is an option to retract or adjust the decision), and iii) whether the consequences would be external or internal to the organisation. Setting up the risk criteria and clearly defining high-risk situations in the early stages of the design process could help to provide interfaces that would facilitate decision-making and reduce risk-related cognitive challenges (Rundmo, 2001). Research shows that individual perception of risk is unreliable, as it often depends on subjective feelings about the technology (Slovic et al., 2005). It also highly varies at an individual level. The way experts perceive the risk of the same situation has been shown to differ between individuals with directly comparable grades of expertise (Williams & Noyes, 2007).

5.4.2.1 Calibrate the perception of risk

A system interface designed to communicate the risk level of the decision being made (e.g., potential consequences or affected individuals) could calibrate the expert's perception of it. For example, a potential explainability interface could include pop-up alerts with additional information regarding involved risks. It could also indicate the situation's severity by using words such as *critical* (Long & Magerko, 2020). Experts show a higher consensus on linguistic risk representations than on numeric ones, thus appropriately selected words could more directly convey the risk (Atoyan et al., 2008). Risk could also be effectively communicated by visualising it or by using indicative colours and symbols (Rayo & Moffatt-Bruce, 2015).

Using design approaches to inform decision-makers about the risk of the decision could also reduce their cognitive load. Experts' cognitive load is increased in high-risk situations because risk evaluation requires tapping into situational awareness and making additional diagnostic decisions (Kaempf et al., 1996). Experts evaluate high-risk situations by matching observed features with their previously learned interpretations of cues and patterns and mentally simulating a story explaining how any situation has occurred (Lipshitz et al., 2001) An explainability interface that orders and visualises features (cues) by their importance (weight) toward the output could help build a coherent story explaining the available evidence and lower the cognitive load (Liebhaber et al., 2002). Contextualising explanations by using visual examples could also ease feature matching and story generation or evaluation and thus reduce the cognitive efforts required from experts (Kaempf et al., 1996). Lastly, using dynamic annotated visualisations instead of simple text-based aids could help to effectively promote comprehension and present the potential risks of different decision scenarios (Rayo & Moffatt-Bruce, 2015).

5.4.2.2 Slow decision-making down and enable analytical deliberation

Decisions made in high-risk contexts, or those with high cost of error, require a slower and more analytical approach. Thus, in increased-risk situations, decision-makers should be slowed down and encouraged to gradually inspect the information provided to avoid fast and heuristic-driven decisioning (Klein et al., 2010). This can be achieved by using interactive explainability interfaces (Cheng et al., 2019), asking experts to acknowledge the explanation or including an extra action needed to access it (Rundo et al., 2020). In high-risk situations, decision-makers can also be supported by allowing them to identify effective options more easily. For example, an interface could let experts compare the outputs to provided prototypes and similar situations (Klein et al., 2010), rule-book protocols or checklists (Dobrow et al., 2006). Having an example to follow has been shown to ease experts' ability to recognise atypical situations that need action or amendments (Klein et al., 2010) and help reduce the related cognitive strain (Orasanu, 2010). Rule-based instructions and schemas can also enable more thorough and analytical data processing (Dobrow et al., 2006; Orasanu, 2010). However, when providing additional resources, one needs to be conscious of the risk of overutilisation bias (the bias to exploit all resources), which happens outside conscious awareness (Reason et al., 1997). Decision-

makers believe they can effectively manage the information and resources available. However, they often do not appreciate the limitations of their ability to regulate the related cognitive workload (Wearing, 2004).

5.4.3 Dynamic 3: Time-pressure

Expert decision-making in both high- and low-risk contexts is moderated by time pressure. Time constraints affect the strategies a decision-maker can employ. Under severe time-pressure, when a slow analytical approach is unlikely, experts are particularly susceptible to heuristic-driven decisioning and various biases. Not addressing this factor in explainability interface design could lead to errors due to decision-makers failing to judge their ability to fill in the gaps in information and make accurate assumptions.

5.4.3.1 Limit the amount of information provided

In high-time-pressure conditions, explainability should aim to reduce cognitive workload. For example, it could be designed to limit the number of alternatives shown at a time, only allowing access to detailed information once a particular alternative is selected (Hutton & Klein, 1999). Providing too much information may be distracting and trigger the utilisation of various heuristics, for example an expert might fail to focus attention on important information (Hutton & Klein, 1999). Providing more information than necessary might also damage experts' performance by subjecting them to overconfidence (Klein et al., 2006b). Moreover, the explainability interface should be designed to provide information gradually. Research suggests that when rapid decisions are needed, providing information sequentially can be more effective, as only one option is being considered at a time (Klein et al., 2010). This way, faster and more intuitive decision-making strategies are supported and cognitive load is reduced.

5.4.3.2 Use visualisations to guide attention

Visualising information can further reduce cognitive load, especially when a large amount of information needs to be processed in a short period of time. It has been shown that experts in time-constrained situations use mental imagery when considering information. This process helps them to recognise cues and visualise possible ways of implementing various solutions and potential outcomes (Klein et al., 2010). For example, visualising information as a diagram can reduce the strain put

on working memory when processing it and speed up the comprehension process (Johnson-Laird, 2010). When information is presented visually, the user does not have to hold and later recover all the information points in their working memory (Johnson-Laird, 2010). Static salient visual features can be used to guide less experienced decision-makers to critical information (Eick & Wills, 1995). In contrast, experts would benefit from being able to freely explore the algorithm through interactive visualisations. For example, by changing the attribute values and observing how the algorithmic decision changes accordingly (Cheng et al., 2019).

5.4.4 Proposed design goals and examples of design strategies

To support the ERT Framework, I propose a list of design goals tailored to the decision-making strategies used under various combinations of expertise and risk dynamics moderated by the level of time-pressure. Each design goal is followed by examples of design strategies that could be used in designing explainability interfaces. The examples without an indication of a specific time-pressure dynamic apply to both high and low levels.

High level of expertise and high level of risk

- **Calibrate the perception of risk.** Use pop-up alerts and/or linguistic indications informing about the risk. Use colors and symbols to visually communicate the level of risk and uncertainty.
- **Facilitate pattern recognition and mental simulation/evaluation of alternative scenarios and help to reduce cognitive load**. Embed explanations in the domain-relevant context. Enable the use of rule-based protocols and checklists, especially in high-time-pressure situations.
- **Support experts' ability to expand information and see *noise* in data**. Allow exploration of multiple variants within categories using refinement tools and clustering techniques.
- **Support engaged and analytical decision-making**. Allow interactive manipulation of attribute values and observe how the output changes accordingly, providing an option to compare and contrast features or categories.

- **Slow down the decision-making process, increase conscious engagement and caution experts against the use of heuristics**. Use an interactive interface and allow decision-makers to actively interrogate the outputs and data. Use design friction and ask experts to acknowledge an explanation or require deliberate action to access it.
- **Support the use of flexible information search strategies.** Provide ways to flexibly explore available information through interactive interface features and adjustable inputs. Allow detailed information to be obtained by selecting a data point or by giving explicit control of which features or categories to compare and contrast.
- **Support serial information processing in high-time-pressure situations.** Provide an option to view a single information point at a time, allowing an easy transition between options.
- **Support the ability to quickly recognise critical information in high-time-pressure situations.** Use visualisations indicating critical information. Provide predictive markers that are highly valuable or highlight weights of contributing features, e.g., use dynamic annotated visualisations.
- **Reduce information clutter in high-time-pressure situations.** Limit the amount of information shown on the interface. Support the expansion of each data point by, e.g., hovering the cursor over it.

High level of expertise and low level of risk

- **Facilitate pattern recognition and mental simulation/evaluation of alternative scenarios and help to reduce cognitive load**. Embed explanations in the domain-relevant context. Enable the use of rule-based protocols and checklists, especially in high-time-pressure situations.
- **Allow flexible information exploration.** Use refinement features to guide the search mechanisms and allow exploration of information in cases of disagreement or uncertainty.
- **Make explanations part of the workflow.** Present explanations in a seamless way, avoid using interrupting features and instead consistently apply visual aspects such as colour-coding and symbols.

- **Reduce information clutter in high-time-pressure situations.** Limit the amount of information shown on the interface. Support the expansion of each data point by, e.g., hovering the cursor over it.

- **Support serial information processing in high-time-pressure situations.** Provide an option to view a single information point at a time, allowing an easy transition between options.

Low level of expertise and high level of risk

- **Calibrate the perception of risk.** Use pop-up alerts and/or linguistic indications informing about the risk. Use colors and symbols to visually communicate the level of risk and uncertainty.
- **Facilitate guided exploration of information in high-time-pressure situations.** Use visualisations indicating critical information and showing the path for information exploration.
- **Reduce cognitive load.** Enable the use of checklists.
- **Support the consideration of time available in high-time-pressure situations**. Adjust the length of explanations depending on how much time is available.
- **Reduce information clutter in high-time-pressure situations**. Provide less detailed explanations but highlight information that is critical. Use bar charts to illustrate the breakdown of the output, including weights of different attributes towards it. Group various attributes by colour.
- **Support analytical evaluation of all the available options in low-time-pressure situations.** Use detailed descriptions of the attributes and features, apply visualisation techniques to make it easier to compare different options. Use dynamic visualisation techniques and colour coding to illustrate each feature's weight towards the output, use interactive refinement tools to show changes in the distributions after updates.

Low level of expertise and low level of risk

- **Support structured information search and facilitate the building of coherent stories in high-time-pressure situations.** Order the features from the most to the least important.
- **Facilitate the guided exploration of information in high-time-pressure situations.** Use visual features indicating critical information and showing the path of information exploration.
- **Support analytical evaluation of all the available options in low-time-pressure situations.** Use detailed descriptions of the attributes and features.
- **Facilitate learning and expertise development.** Allow interactive questioning of the output and provide feedback options.

5.5 Speculative scenario of the ERT Framework application in a journalism domain

Following the development of the ERT framework, a demonstrative case study in the context of journalism was developed to illustrate the potential application of the ERT framework in practice. Two speculative scenarios were developed in collaboration with an expert journalist and researcher who was consulting with her experienced colleagues embedded within the media sector. These scenarios contained narratives of how journalists could be reasonably expected to engage with two different decision-support systems and how a designer or a researcher observing them could draw from the ERT framework to inform their XAI strategies. This part of the thesis was prepared in collaboration with an experienced journalist and researcher, Dr Bronwyn Jones, for the publication *Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable* (Simkute et al., 2021).

5.5.1 Methodology

A speculative scenario-based design method was chosen to help anticipate and leverage scenarios of possible framework use at an early stage of system development and to understand unanticipated drawbacks or new requirements when XAI systems are deployed into complex use settings (Wolf, 2019). Using abductive reasoning, scenarios were developed by working in collaboration with an experienced journalist and researcher who was consulting with her colleagues working in the media sector. This method helped to test the potential utility of the

framework and elicit aspects of the problem space AI opens in news production. The scenarios were also informed using secondary research sources from the media domain (Diakopoulos, 2020). Scenarios were developed to align the potential applications of the ERT framework with the existing use cases of AI in a media context. They were also developed speculating how those use cases may be changed when a new AI system is introduced (Wolf, 2019). The plausibility of the scenarios was tested by an expert responsible for developing and testing newsroom technology and several experienced journalists working at the BBC. Their feedback regarding context, probability, and coherence was incorporated into the scenarios. Research suggests that plausibility is a central criterion for validating scenarios as representational products and ensuring their heuristic effectiveness, making it reasonable to believe that they could happen in real-world situations and that they are trustworthy and credible (Urueña, 2019).

This part of the study aims to answer the question: how can the ERT Framework be applied in real-life scenarios? The scenario-based design approach provides narrative descriptions of envisaged usage episodes (Wolf, 2019) that could help to guide the development of a system interface (Sears & Jacko, 2009). Here, I use scenarios as examples of when journalists interacting with AI could require explanations and how designers might deploy the ERT framework when designing explainability interfaces.

5.5.2 The AI-driven systems in the journalism domain

AI-driven systems are gaining traction in journalism, where journalists increasingly rely on them to gather, produce, and distribute (Diakopoulos, 2019; Marconi, 2020). Recommender engines drive new forms of audience personalisation and engagement (Helberger et al., 2020), audience analytics tools drive subscription and monetisation strategies, and semi-automated content production systems generate stories, visualisations, etc., with little human intervention. Crucially, decision-support tools now underpin elements of editorial decision-making in the newsroom, such as text and image classification and suggestions, data analysis, and media monitoring. They have the potential to bring time and resource efficiencies (Marconi, 2020), opportunities for wider oversight (Diakopoulos, 2019), deeper analysis (Stray, 2021) and greater creativity (Maiden, 2018) but also risk disrupting long-established ways of working. Newsworkers consider ‘journalistic intuition’ and ‘gut instinct’ fundamental

to their job. Specialised knowledge and discretion are central to journalistic self-conception (Christin, 2020). However, there is evidence that newsroom culture has shifted following the algorithmic intervention, for instance, toward placing more value on analytics than professional intuition (Hanusch, 2019). Most journalists have little understanding of how these systems work and limited ability to critically assess automated outputs or their suitability in context (Jones & Luger, 2021). This knowledge and communication gap risks leading to journalistic malpractice (Hansen et al., 2017) and undermining public confidence in ethical and responsible journalism. If this is to be avoided as AI systems become more pervasive, there will be a need for explainable interfaces that account for the demands of the journalistic context. Despite this, there has been little focus on explainability in journalism and seemingly low levels of recognition amongst news organisations that this issue needs tackling.

The core of a journalist's job is creating news content promptly by making decisions efficiently and exercising expert judgement within a framework of laws, regulations, professional norms and socio-cultural expectations. In contrast to many of the high-stakes areas often prioritised in explainability research where the risk profile is often mediate and extreme (e.g., life and death decisions in defence, medicine, etc.), the risk of opaque AI systems in news production is one of the aggregated errors, unrecognised biases and cumulative oversights. This can lead to inaccuracies and the inability to sufficiently account for and justify editorial decisions, which can harm news organisations' reputations and undermine the legitimacy of journalism in society. Knowledge claims in journalism are subject to varying criteria for adequate justification, which complicates the task of communicating the role of AI systems in decision-making to both editors and the audience. Diakopoulos (2019) highlights the importance of building smart interfaces to support journalists, including designing suitable signals to highlight relevance and other indicators of newsworthiness, as well as engendering appropriate trust by reflecting uncertainty. This can include "multi-modal and interactive interfaces", "summary sentences of text", providing an explanation, or "visual evidence and context" (Diakopoulos, 2019, p. 82). The ERT framework can contribute to responding to these challenges.

Scenario 1: Image suggestion decision-support tool

Leila is part of a team that is designing an explanation interface for an image suggestion system that uses AI for facial recognition and image quality classification. Journalists will use the system to help identify and choose the best images quickly from a wide selection. They need to justify to their editor why they chose the images and trust that the system's suggestions are accurate and appropriate. Leila wants to understand which type of explanation will be most useful for the journalist-user, so she observes several who are writing stories for the website as they use a system prototype, which does not yet include an explanation component. While observing, she maps each scenario onto the ERT framework, asking herself: what is the level of expertise, risk and time-pressure here, and what does the framework recommend as an appropriate explanation approach?

The first journalist, Ada, is writing a breaking news story about the meeting of political leaders for a G7 summit in Cornwall and is expected to get the story live within minutes. An experienced journalist, she is comfortable making editorial decisions about which images to use. Still, she is not up-to-date with global political leaders, and this is the day's top story that millions of people will read on the website homepage, so any errors could cause reputational damage to the news organisation and herself. As Ada types about a meeting between UK Prime Minister Boris Johnson, US President Joe Biden and South African President Cyril Ramaphosa, the decision-support system identifies keywords to suggest a 'top 5′ selection of images based on relevance and image quality. Ada rapidly scans the images – she is unsure what Ramaphosa looks like, so she chooses the top-rated picture of what she believes is the three men. Ada wishes she knew why the system was rating this one so highly and whether it is sure the third man is Ramaphosa, but no explanation has been provided. Ada is used to search for relevant images in picture libraries. Still, she has never used this type of AI system, which recommends a selection automatically, so she feels unsure how accurate it is and how much trust to place in the algorithms. Her wariness prompts her to call over a colleague in the newsroom to check the image, and she searches online for pictures of the South African leader to compare. This cross-checking reveals that the image is of Johnson, Biden and an unknown man – so she removes it from the story and scrolls through the recommended options until she finds a suitable replacement that she can

corroborate as being the correct leader. Once submitted for sub-editing, she turns to her colleagues to discuss why they think the AI system made a mistake and whether the error might slip past the attention of a less experienced journalist.

Leila refers to the ERT framework and characterises the situation: high journalistic expertise but low topic expertise, high-time-pressure and high risk of reputational damage. For this combination of factors, the framework suggests that enabling a more analytical approach and preventing the “illusion of validity” (e.g. by providing a rule-book protocol to match and compare her situation against) could help in a higher-risk situation like this. Emphasising the most important aspects to support quick and intuitive pattern recognition, for example, by showing feature weights, could be useful for someone with high expertise. Finally, in such a time-pressured situation, moderating the amount of information provided, ensuring it is presented sequentially, supporting mental simulation/imagery and reducing cognitive strain by visualising information.

Using the ERT tool, Leila sees that the goals of explainability interface design in this situation should be to a) support serial information processing, b) reduce information clutter, c) calibrate the perception of risk, d) facilitate pattern recognition/reduce cognitive load, e) support the ability to quickly recognise critical information, and f) slow down the decision-making process. Guided by the ERT framework, Leila decides to use an interface design which would allow Ada to view a single information point at the time and would have a feature allowing her to easily transition to the next information point (support serial information processing) that would limit the amount of information provided, by would allow to hover the cursor over information points and expand them (reduce information clutter), the interface would also use visual features to indicate critical information and highly valuable factors embedding them in the context (support the ability to quickly recognise critical information/support pattern recognition/reduce cognitive load) and would inform Ada about the level of risk (calibrate the perception of risk).

The second journalist she turns to, Marc, pulls together a round-up photo gallery of the best images from the Cannes Film Festival. He is new to the job and has no experience working with imagery or entertainment coverage, so he is happy that the system can help him filter through the hundreds of pictures on the system. There is no strict deadline, so he can take his time. Marc uses his judgement of what makes a good picture coupled with what seems newsworthy, seeking out big-name stars,

surprise winners, and out-of-the-ordinary happenings. Still, he also allows the tool's suggestions to help guide him and clicks through its recommendations as they appear. As he types general search terms, including the festival name, the name of its prizes, and certain celebrities, the system generates recommended images with metadata attached, including the photographer's description, date taken and copyright information. As he does this, he finds himself questioning why each image has been picked, as he is not always clear on the connection between the terms he has used and the image or how he can assess the accuracy of each suggestion. To double-check, he searches online for stories about the celebrities depicted in the images to get a better sense of their relevance and importance and writes captions for those he chooses to include. Leila notes that this situation has low journalist and topic expertise, low-time-pressure and low risk of reputational damage. Because there is low risk, explaining (or visualising) why some of the features are more salient than others, providing explanations/information in a neutral manner (neither negative nor positive) could help to prevent a framing effect (leaning to certain decisions due to the way information is framed), which is especially likely in low-risk conditions (Tversky & Kahneman, 1974). As Marc has low expertise, providing more information and supporting guided comparison and evaluation of the features (e.g., showing feature weights with accompanying explanations/suggestions) may enable practice, learning and potential development of expert skills. Because he is under low-time-pressure, it is advisable to enable a slower and more deliberate analysis of information by allowing (and encouraging) Marc to question and challenge predictions and investigate the importance of features (e.g., by being able to ask questions, interactively communicate with the system).

Leila uses the ERT Framework and decides that the goals of explainability interface design in this situation should be to a) support analytical evaluation of all the available options and b) facilitate the development of expert skills. Leila decided to use an interface design which would allow Marc to see detailed descriptions of the attributes and features, apply visualisation techniques to make it easier to compare different options and use dynamic visualisation and colour coding to illustrate each feature's weight towards the output (support analytical evaluation). The interface should also be interactive, allowing Marc to further question the output (facilitating learning).

The following day, a colleague sends Marc a blog post critiquing how his news organisation has “erased black and minority ethnic winners at Cannes” by prioritising images of white celebrities. Reflecting on his work process, he realises that only white celebrities were recommended by the system and wonders how and to what extent the system’s recommendations impacted his decisions about which images to include.

Scenario 2: Data-mining and visualisation system

Interface designer Jo has been tasked with considering how to build explanations into a new AI-driven data mining and visualisation tool for investigative journalists that finds and displays connections between data. She is sitting with Salim, describing his work process as he conducts background research on a story. Salim says: “I’m digging into the background of a well-known politician to find out more about his business dealings, and the tool has pulled together this visual map that shows links to publicly available documentation that mentions him. See here (he points). The image shows clickable nodes denoting the person or company or ‘thing’ he’s mentioned alongside links to the source of the info. At first glance, it seems to suggest he’s tied to more than 20 offshore companies and several criminal figures.” Jo asks if he understands how the system made these connections and if he trusts it. As he clicks on various links to published articles, public records, and data sets to explore further, he says: “I just treat these like tip-offs or suggestions of things I might want to look into. They might come to something after I check them out, or they might lead nowhere, but I don’t take it as given that what the system suggests is right. I’ve still got to do all the hard work checking out all of these leads and seeing what I can stand up and verify.” Salim thinks for a moment and is silent before adding: “To be honest, I don’t know how it works really... I guess the AI is crawling the web and finding things and making links between things I haven’t seen before. Still, if a libel suit comes in, I could hardly use the defence: ‘The AI did it!’” Using the ERT tool, Jo assesses this to be a high-risk, high expertise, and low-time-pressure scenario. Though Salim made it clear that he was using the tool solely as a stimulus to point him in the direction of potentially interesting information or highlight connections he might not have made, Jo notes that he pointed out the risks associated with the type of investigative work the tool is designed to support. The high risk of making poorly substantiated claims leading to legal action and

reputational damage suggests that allowing Salim to actively question the data and providing more information would be a good explanation strategy. Experts in high-risk, low-time-pressure situations like these are highly motivated and more likely to challenge their intuitions and explore more information (Svenson, 1979). High expertise suggests it would be beneficial to allow information search in an expert's chosen way (e.g., by providing refinement tools) and providing access to 'raw' and 'noisy' data. Because this is a long-term project and Salim is under low-time-pressure, supporting slow analytical thinking would be beneficial by allowing him to manipulate the data and make comparisons (e.g., interactive model simulations).

Using the ERT tool, Jo sees that explainability interface design goals in this situation (high risk, high expertise, and low-time-pressure) should be to a) slow down the decision-making process, b) support flex information search strategies, c) calibrate the perception of risk, d) facilitate pattern recognition/reduce cognitive load and e) support the ability to see *noise* in data. Jo decided that using an interactive explainability interface where Salim would have to actively engage with the explanation would be suitable. This could be achieved, for example, by having to take an extra step to access it (slowing down the decision-making process), being able to manipulate attribute values and observe how the output changes accordingly (support slow analytical deliberation), accessing additional information by selecting any data point/feature (flexible information search) and being given the option to explore multiple variants within categories using refinement tools (exploration of *noise* in data).

5.6 Limitations of the ERT framework

It is essential to recognise the pressures facing organisations wishing to develop their explainability interfaces, as these real-world factors can surface challenges and sometimes insurmountable constraints to applying frameworks such as the ERT. Time pressures and gaps in expertise can hinder even those design and development teams with the best intentions. Though the framework offers a *shorthand* for considering pertinent insights from Cognitive Psychology and HFE, any team using it would need to allocate time to scoping, in advance, the expertise, risk and time-pressure profile (s) of potential users. The ERT framework would also likely be one of several tools needed for any holistic analysis of the optimal explainability approaches.

By replicating human decision-making strategies, we should also be careful not to transfer human biases into algorithm-supported decisions. Although expert decision-making has unique qualities, all humans are susceptible to using various imperfect heuristics. In the review, several of these heuristics were touched on briefly, suggesting how explainability could be used to avoid them. However, more research should be done to examine how enabling naturalising decision-making affects the transfer of these and other types of biases and heuristics.

The speculative journalism scenarios illustrate how the ERT tool can guide designers towards evidence-based assessments of the optimal approaches to planning explanation interfaces that do not flatten out explanations in a way that suggests a single approach is adequate. However, it also indicates a limit to what can be achieved with the tool when situational profiles are dynamic. However, the scenarios demonstrated that the ERT Framework could be valuable for guiding interface designers. Moreover, these scenarios showed another important aspect of the ERT Framework. It makes the interface designer or another responsible stakeholder more mindful of who the intended stakeholder is and in what context they make decisions.

5.7 Interface design guidelines for experts in high-risk contexts

After developing the ERT framework and speculative cases of its application, I aimed to expand the framework and match suggested design goals with concrete interface design solutions. Workshops with interface designers at various stages of their professional careers were organised to map design ideas to the key design goals supporting expert decision-making and reasoning strategies in high-risk and high-time-pressure environments. The workshop participants combined their own experience with the new knowledge of the ERT framework to propose various concrete explainability interface design solutions. These solutions were grouped into several broader and non-restrictive categories. Although the proposed solutions could be applied as provided by selecting one or multiple solutions for an XAI project, they are not prescriptive. They can simply be used as a source of information. Hypothetically, an interface designer or explainability researcher could select the solution from the list or map their own design solutions that fit within suggested broader categories. These interface design guidelines can also serve as a brainstorming tool, allowing experts to be more mindful of expert decision-making

and reasoning limitations and capabilities while developing novel interface design solutions. The proposed guidelines are domain-agnostic and generalisable. All the provided solutions can be tailored to be domain- and task-specific by the interface designers and XAI researchers. To my knowledge, the ERT is the first XAI design guideline developed to support domain experts. They are also the first XAI design guidelines emphasising the importance of high-risk and time-pressure situations and proposing ways to address increased cognitive load in these contextual circumstances when designing explainability interfaces.

5.7.1 Methodology

Four design workshops were organised to explore how explainability interface design could be tailored based on the general guidelines provided by the ERT Framework. The key goal of these workshops was to distil concrete examples of interface design features that would fit within the proposed ERT design goals for high-expertise, high-risk cases in both low- and high-time-pressure contexts. The ERT suggested design goals were:

1. **To uncover cues and patterns and enable mental simulation of features.** To support the exploration and evaluation of several possible outcomes and to provide the option to see *noise* in data.
2. **To support flexible information search.** To allow users to select which information and in what order to explore.
3. **To calibrate the perception of risk.** To inform the user about the risks involved, to show the level of uncertainty and/or specific risks involved in the decision.
4. **To slow decision-making down and enable analytical deliberation.** To encourage the decision-maker to actively engage with the explanation and to support a more thorough analysis.

The high expertise dynamic was selected due to the thesis being focused on explainability for domain experts. The high-risk dynamic was established based on i) the explainability literature review suggesting the high importance of explainability support in high-risk contexts due to the increased cost of errors and the importance of the human-in-the-loop role and ii) the HFE literature review suggesting that experts apply specific sensemaking and decision-making strategies in high-risk

situations and are then more likely to succumb to heuristic-informed decisioning. Both high- and low-time-pressure cases were explored due to their ability to moderate decision-making in high-risk scenarios.

5.7.1.1 The research aims and questions

First, I aimed to improve the ERT framework by recognising aspects that might be challenging for designers and to aid the understanding of how the ERT framework could be made more usable for design practitioners and decision-makers. To address this aim, I raised the following research questions:

- *What aspects of the ERT framework are seen as the most useful by designers?*
- *What aspects of the ERT framework do designers struggle to envision applying in practice?*
- *Can designers recognise risk, time, and expertise dynamics? How could these dynamics be redefined to fit real-world scenarios from the designers' perspective?*
- *Are there any extra dynamics that should be considered besides the risk, time and expertise?*

Second, I aimed to match the proposed design goals with concrete design approaches in order to populate the ERT framework with examples of design strategies and shape informative design guidelines for explainability interface design. To address this aim, I raised the following research questions:

- *How do designers envision the suggested design strategies from the UX perspective?*
- *Which design strategies do designers see as the most important?*
- *Which design strategies do designers see as the least important?*
- *What concrete examples of interface design features do designers match with suggested design strategies?*

5.7.1.2 Participants

In total, 39 participants attended the virtual workshops. Seven postgraduate Design students from the University of Edinburgh participated in the pilot workshop. They were recruited through the invitation email distributed via the Design Informatics

Department mailing list. Postgraduate students who expressed their interest in the study were sent an online consent form and information sheet via email at least three days before the workshop (Appendix D). Participants of the pilot workshop received £10 bookshop vouchers as compensation for their time. The pilot workshop was conducted to help improve the workshop procedure and to gather participants' feedback about the workshop. Given that the BBC employees were participating during their working hours, I aimed to avoid time-costly procedures and only conducted a more concise final version of the workshop with them.

The following three workshops included 32 UX designers working within three different departments across the BBC. Participants' areas of expertise varied from news, innovation, product design and interface design research. Participants were recruited using a snowballing effect by contacting leads and managers of various UX departments within the BBC and asking them to inform potential participants about this workshop so they could then contact the principal researcher. All participants received information sheets and consent forms (Appendix E) via email at least three days before the workshop and either signed a consent form or gave a spoken consent statement during the Zoom recording. The BBC employees participated in the workshops during their working hours and were not compensated for their time due to the internal BBC rules.

5.7.1.3 Procedure

The pilot workshop, which took 90 minutes to complete, was conducted via Zoom video call. First, workshop participants were given a brief presentation explaining the workshop's aims and giving an overview of the ERT Framework. Then, participants were invited to comment on the ERT Framework and ask questions about it, aiming to see the key usability issues of the framework and aspects that could be unclear to the designers using it in real-life situations. Participants then received a Miro board link and were guided through the journalism scenario (high-risk, high-time-pressure, high-expertise) that covered both journalists' and designers' perspectives. Scenarios can be found in Appendix F. Participants were asked to read the journalism case scenario and recognise whether the situation was high- or low-risk, high- or low-time-pressure dynamics, and high- or low-expertise. This was done to see if they understood and could recognise the dynamics. Participants were asked to critically evaluate the real-life applicability of the framework and suggest missing dynamics

and roadblocks that could arise during the implementation of the framework. Then, they were shown and explained five design goals and were invited to ask clarifying questions. Afterwards, they were given time to individually brainstorm design features to match the provided design guidelines from the ERT Framework using virtual sticky notes assigned to each goal. The ERT Framework, design goals, and scenarios were visible to participants during the task on the same Miro board as they were working on during the task. Screenshots of the Miro board are included in Appendix G. Then, participants were invited to enable their microphones and contextualise, reason, explain, and discuss their design suggestions. Afterwards, participants were invited to complete the same task using a social work scenario (high-risk, low-time-pressure, high-expertise) (Appendix F). Participants were given 35 minutes to complete both parts, including the group discussion. After the workshop, participants said that it was challenging to attend to both scenarios. Reflecting on this feedback, the following three workshops were shortened to 60 minutes but only had a single journalism scenario (high-risk, high-expertise, high-time-pressure). However, the outputs of the pilot workshop were still useful, informative and aligned with the responses of the UX designers, so they were included in the data analysis and final report.

The following three workshops were conducted with the BBC UX designers and took 60 minutes to complete. They were also conducted via Zoom video call. The workshop followed a procedure similar to that of the pilot workshop, but it only included a journalism scenario. Instead of the introductory presentation, participants received a briefing on the workshop goals and procedure, as well as an explanation of the ERT Framework. All the information was included in the Miro board. Screenshots of the Miro board are included in Appendix G. Participants were given 35-40 minutes to individually complete the main task and 10-15 minutes to explain their suggestions in a group discussion. The feedback about the framework was left to the end of the workshop and was given around five minutes if there was time left after the main task completion.

5.7.1.4 Scenario generation

The journalism scenario was an adapted version created for the speculative scenarios discussed in the previous subsection *Speculative scenario of the ERT Framework application in a journalism domain*. The social work scenario used in the

pilot study was based on the speculative application of the AI prototype to process asylum and refugee status applications (Ahmad, 2020).

The scenarios were intended to portray the story of an expert who uses the technology and needs an explanation and a designer who applies the ERT Framework to design an explainability interface. This strategy was chosen to put the participants in both the designers' and users' shoes so they could judge what information is needed to understand which dynamics and their levels are most prominent in each context. This was done with the intention of helping participants evaluate the ERT Framework's applicability from the designers' perspective and empathise with the experts' needs. Scenarios are included in Appendix F.

5.7.1.5 Collected data and measures

During the pilot workshop, the Zoom video call was recorded, and notes were taken throughout. During the workshops with the BBC employees, only digital notes were taken, as recordings were not allowed due to internal BBC rules. Notes were taken to capture participant explanations of their design suggestions, feedback about the framework, and other relevant responses to the study aims and questions. Participants' design ideas were captured as digital sticky notes on the Miro board during all four workshops. These notes were screenshotted and saved. Only the data from the Miro boards and notes were used for this study. Any identifiable and confidential information was edited to remove personally recognisable information. All participants were informed about their right to withdraw from the study at any time and for any reason and have their data removed from the study with no penalty or loss of benefits to which they may be otherwise entitled. The Edinburgh College of Art Ethics Board approved the study. More details about the study data handling, access, storage and privacy can be found in Appendix H.

5.7.1.6 Data analysis

The workshop data analysis was informed by Grounded Theory (Glaser and Strauss, 1967). First, the design suggestions were collected from each Miro board, and then the notes from the workshop were reviewed for participants' elaborations on the suggestions. The data was separately analysed for each design goal. However, during the workshop, most participants expressed that the design goals "Uncover cues and patterns" and "Enable mental simulation of features" were too similar. The design suggestions for these two goals overlapped and repeated. Thus, the two

goals were combined, and their design suggestions were grouped before the analysis. The suggestions were categorised using inductive analysis and searching for overarching themes, with various examples of UX design features falling under several categories. Some of the examples were marked by participants as suitable for multiple design guidelines and were kept in multiple areas. However, different benefits of the suggested features for expert decision-making were emphasised, and a different application angle for each suggestion was proposed. The time-pressure was used as a moderating factor, and suggested design features were grouped into suitable for high-, low-, or both time-pressure situations. The applicability depends on how time-consuming the task was, the amount of information it required for the user to attend, whether it supported sequential processing, and whether it was reducing mental workload.

5.7.2 Design suggestions for the ERT framework guidelines

Design Goal 1: To uncover cues and patterns and enable mental simulation of features – to support the exploration and evaluation of several possible outcomes and to provide the option to see noise in data

Uncovering cues and patterns by providing relevant contextual information could improve the information analysis and trigger recognition of errors or other irregularities. Expert analysis and the ability to effectively monitor automation suffer when outputs are stripped of supporting information (Stuart et al., 2007). Experts can better apply their expert knowledge when they notice aspects that evoke their prior experiences (Hutton & Klein, 1999). They also perform better if they can mentally manipulate various features to find the most appropriate solution (Hutton & Klein, 1999). Using design features that could enable both pattern recognition and simulation of features could support expert decision-making, enable their skill application, facilitate their ability to recognise errors, and reduce overall mental workload. Workshop participants suggested various interface features that were grouped into five broader categories.

Categorisation

The explainability interface could allow the categorisation of information, emphasising and directing the expert's attention to the information that could be

potential cues and patterns. Categorisation could be particularly useful in high-time-pressure situations where the cognitive load should be reduced. It could help experts find the information they need more efficiently and support sequential data processing. Categorisation could also enable a more structured way of assessing information, allowing deeper information analysis.

- **Themes.** The XAI interface could organise information thematically by using keywords to determine the main themes. To increase user agency, experts could edit the themes to adjust them to their specific needs and tasks.
- **Tags.** The XAI interface could use tags to categorise explanatory or supporting information. Tags would inform the user about the categories to which each item belongs.
- **Ranking.** The XAI interface could use a hierarchical order to display information, reveal *quick finds*, and provide more detail if a specific information point is selected. The XAI interface could also allow the user to pre-rank factors/features so that the most important elements are displayed when a lot of data/information is shown.
- **Filtering.** The XAI interface could provide an information filtering feature, allowing an expert to find and filter out *similar* results. Filtering could enable a navigation system, allowing users to *jump in* and *out* of the current state.

Comparing

The explainability interface could support pattern recognition and deeper data analysis by using features that would allow for comparing several factors and enable cross-referencing of different concepts. This approach could support quick recognition of the factors that differentiate cases or make them similar, thus triggering relevant experiences that could help experts apply their prior expertise.

- **Show comparison.** The XAI interface could allow users to access a library with a combination of different outputs or metadata to compare these results. The interface could enable users to compare aspects they see as important in each situation, such as features, tables or cases.
- **Similarities and differences.** The XAI interface could group the information based on similarities and differences. The interface should make it easy for the user to quickly see similar results in one group and different in another,

thus making it easy to compare and notice irregularities. The XAI interface could display adjacent concepts and information points.

- **Criteria.** The XAI interface could allow the user to set the criteria based on which the results/data should be displayed. If no criteria exist, the results could be displayed based on the most significant factor, e.g., confidence or best match.

Highlighting

The explainability interface could include features highlighting important information for a decision and supporting experts' pattern recognition and mental stimulation. Highlighting is intended to draw experts' attention to the most important factors (with high confidence that they are accurate and important), reducing the cognitive load that experts experience in high-risk situations.

- **Key parameters.** The XAI interface could highlight the key parameters used to generate the output or the features that had the most weight in determining the final output.
- **Confidence.** The XAI interface could highlight when the confidence in the output is low or below a set confidence threshold. The same highlighting function could then emphasise potential reasons for the low certainty or draw attention to specific areas.
- **Commonalities and variations.** The XAI interface could highlight common aspects between the features of outputs, for example, by highlighting the exact weights of the contributing elements. It could also highlight aspects that vary, for example, by highlighting variation across the results/data points/outputs.
- **Outliers.** The XAI interface could display and highlight factors and data points outside the usual range and emphasise the information about outliers and inconsistencies in the output.

Gamifying features

The explainability interface could include additional features that would gamify the process of comparing and analysing results, allowing the user to engage and interact with the system in a more structured but stimulating way.

- **Lightbox.** The XAI interface could include a lightbox to save searches and matches with a quick view of confidence scores.
- **Playground page.** The XAI interface could have a playground page with sliders of various features to help the user understand how these features interact and affect the outcome.
- **Mix & match.** The XAI interface could have a feature that would allow mixing and matching features, enabling the exploration of different pairings of the results/outputs.
- **Persistence.** The XAI interface could have a feature that would refresh the results when they do not match the user's expectations or are recognised as incorrect.
- **Parallel searches.** The XAI interface could include a feature that would allow proceeding with multiple searches or filtering of the results and then would allow the comparison of these search suggestions.

Examples

The explainability interface could include examples of similar situations to help experts recognise patterns and simulate various scenarios more easily. They could also compare the results to the examples, which could help them to spot potential irregularities. Depending on the context, experts could see real-life examples or examples from previous searches to illustrate the logic behind the model's working and help to notice irregularities in a newly generated output.

- **Real examples.** The XAI interface could show a real example of the model getting it wrong or right, helping the user see any similarities and differences to the case in question, e.g., showing an example of an incorrect image with a high confidence score.
- **Case studies.** The XAI interface could link case studies illustrating situations when things go wrong, highlighting similarities and differences with the current case.
- **Additional learning resources.** The XAI interface could link relevant learning resources with multiple examples that would help the user to familiarise themselves with similar and different cases and build a more in-depth understanding of the situation and how an output fits within it.

Design goal 2: To support flexible information search - to allow users to select which information and in what order to explore

Allowing experts to investigate and question the output interactively and flexibly could support their preferred information search methods. Experts use unstructured search strategies and feel restricted when they cannot do that (Klein et al., 2006a). Thus, an explainability interface should be designed to support experts' need for flexibility in accessing data, and it should also enable their active involvement in information exploration. Workshop participants suggested several ways to enable and enhance information search. Their suggestions were grouped into three categories.

Filtering

The explainability interface could have an information filtering feature, providing a flexible but organised way to search for information. This approach would allow them to filter search results and find the information experts seek when they decide they need it.

- **Search bar.** The XAI interface could have an instant filtering feature, such as a search bar, to filter the results based on a particular prompt. It could allow users to select search terms for filtering information and click on the selected factors, data points, outputs, etc. These selections could then be fed to the ML training data.
- **Checkboxes.** The XAI interface could enable various checkboxes to guide the filtered search, allowing the user to select multiple search items, search criteria, or search pathways.
- **Breakdown filtering**. The XAI interface could simplify the search by breaking the results into categories or combinations of factors (this would depend on the context). It could allow the user to see a brief breakdown first, enabling them to quickly spot the category where their required information could be found.

Tags

The explainability interface could enable a flexible search, allowing a user to tag and prioritise items or use a certain referencing or logging system to provide a systematic

view of how items are shown in the search. This could aid the recognition of the best way to find information. Using tags could help users find the information faster and more systematically, which could be useful in high-time-pressure situations. Tags could also help to explore available metadata and data groups.

- **Referencing.** The XAI interface could have a referencing system, which would be developed with experts' input. The referencing systems could be based on user feedback and be edited and continuously improved by allowing users to add, edit or delete references.
- **Logging.** The XAI interface could use a logging system allowing users to log different search pathways and make the search more systematic**.**

Prioritisation

To enable a flexible search, the interface could allow users to prioritise search terms and to actively edit this prioritisation depending on their changing search goals and preferences.

- **Prioritise search terms.** The XAI interface could allow users to prioritise search terms to see the information they think is the most important or should be set at the top. Users could edit this selection, depending on what is expected in the search results.
- **Drag & drop.** The XAI interface should allow the user to drag and drop the items, depending on the way they prefer/expect the prioritisation/search to be conducted. This feature would allow users to drag search terms into the desired order.

Design Goal 3: To calibrate the perception of risk – to inform the user about the risks involved, to show the level of uncertainty and/or specific risks involved in the decision

Calibrating the perception of risk could help experts quickly spot the risk indications and reduce their cognitive workload by directing their attention to the key information that could allow them to identify the level of uncertainty or particularly risky factors (e.g., low certainty points or sensitive data, such as gender, or race). AI strips the outputs of informative contextual information, making it difficult for them to challenge their intuition or AI-provided information (Billings, 1991; Lee & Seppelt, 2009; Moray et al., 1986). In high-risk situations, experts tend to search for ways to reduce

uncertainty by comparing their previously experienced situations to the newly encountered ones (Orasanu, 2010). Guiding their attention to the specific points that determined the output could allow them to link each case with a similar one they had experienced before and spot irregularities that might indicate an error. The workshop participants' interface design suggestions were grouped into five wider categories.

Use coding to indicate a level of risk or uncertainty

The explainability interface could inform the user about risk factors such as a high uncertainty of the prediction, a new/relatively untested algorithm being used, a history of unreliable performance, or novel situations that the algorithm might not be trained to deal with by assigning a clearly defined code (e.g., colour) to various levels of these aspects. This would allow experts to quickly calibrate the perception of the factors they should be cautious about. Using code could be particularly useful in high-time-pressure situations, where experts need to be able to spot important features quickly.

- **Colour.** The XAI interface could use colours to identify risk and reliability levels of the output or system. Colour codes could be selected based on the colour theory, considering users' potential biases towards various colours. The colour code could also be based on the traffic light system, i.e., the highest risk would be indicated by red, the average risk by yellow, and the low risk by green colours.
- **Typography.** The XAI interface could use typography to indicate the levels of risk and reliability of the output and system. Different risk and reliability levels could be indicated using typography code, such as a specific font (bold letters).
- **Heat scale.** The XAI interface could identify the risk and reliability levels using a heat scale. Each level could be associated with a specific temperature. For example, the highest risk would be indicated by a hundred degrees Celsius and the lowest by zero degrees Celsius.

Using advanced visualisations

In certain contexts, data visualisation techniques could calibrate risk perception. This is particularly relevant in domains where data analysis is one of the experts' tasks.

Data visualisation techniques could help calibrate the user's perception of risk or reliability, allowing them to connect risk to the relevant data.
This approach is appropriate in high-time-pressure situations when experts have sufficient skills to interpret visualisation with relative ease.

- **Simple visualisations**. The XAI interface could use bar charts and other simple visualisations, allowing experts to quickly recognise risks without feeling overwhelmed.
- **Complex visualisations**. The XAI interface could use heat maps and other complex visualisations, allowing experts to generate a deeper/contextualised understanding of related risks.
- **Interactive visualisations.** The XAI interface could also illustrate and contextualise multiple risks based on different decisions and analyse various options flexibly.
- **Original visualisations.** The XAI interface could illustrate risks, reliability, and uncertainty using original visualisations that are domain-specific and are potentially designed in collaboration with experts. For example, visualisations could have a navigation-like interface. The map could outline and link features and risks to the output, allowing users to navigate the map and explore relationships and links between these aspects.

Using percentage mark

Risk levels could be made visible to the expert user by adding percentage indication marks that explicitly show how likely the output is correct or the likelihood of undesirable outcomes or errors. This approach is suitable in high-time-pressure circumstances when the user needs to be aware of the risk. However, it could be more beneficial under low-time-pressure, where users would have time to inspect the percentage and explore the explainer labels.

- **Percentage with an explainer.** A risk percentage and an explainer label could accompany the XAI interface. Depending on the context, the explainer could provide information about the meaning of the risk and/or what this percentage could mean in terms of risk for the specific decision.

- **Percentage tag**. The XAI interface could provide a tag, such as the likelihood that this output is accurate. The tag could provide additional information about this label if a user requires it.
- **Confidence/certainty percentage.** The XAI interface could use percentages to indicate the confidence or certainty of the output and how conclusive or complete the dataset is. This design feature could show experts if the decision needs further attention and in-depth examination and if the output is recommended to be overturned.

Use indications of verification

Indications of verification could be used to show trustworthy and highly reliable aspects (e.g., outputs or data points) and flag aspects that should be explored further or challenged if not verified. The source could also be verified or marked as trusted, unknown, limited, or risky. Verification could also indicate how well the model performed in the past; this metric could be human-verified and based on the verification process during the low-time-pressure cases. Confirmed/verified cases could be recorded/tracked, and an appropriate verification label could be assigned to the next user. Identifying verification can be helpful in high-time-pressure situations, as it could be a quick way to reduce the amount of information that needs further analysis. It could also reduce a user's cognitive load by showing them aspects that do not need their attention.

- **Mark of verification.** The XAI interface could use a mark of verification, highlighting trusted, complete, or often-used sources. Verification could indicate that something has been previously used or selected multiple times by other users (even showing who used it).
- **Flag unverified information.** The XAI interface could use a specific warning icon to flag when the source/output/algorithm is not verified as trusted or complete. When something is flagged as untrusted or unverified, the user should be required to examine the case further.
- **Feedback through verification.** The XAI interface could have an additional feature allowing the user to apply their verification in cases where they assessed the case and proceeded with it. In low-time-pressure situations,

users could be asked to assess the risk of unverified cases, apply their risk score, and annotate it.

Add context about the extended risk of the decision

Explanations could be embedded in a context that would illustrate potential risks related to a decision. This could help users to be aware of the bigger picture and, in the case of expertise, could provide additional information that could help link a specific case to prior experience and spot irregularities that should be further inspected. This approach could be helpful in both high- and low-risk situations, but the amount of additional information should be limited in the high-risk contexts.

- **Supplementary material.** The XAI interface could include a panel with information about the output, algorithm, or sensitive data. For example, supplementary metrics could be used to show how the model performs on different ethnic backgrounds.
- **Recontextualise.** The XAI interface could use features that recontextualise the output (i.e., embed it in a different context) and aid confidence with decision-relevant examples, such as “94% confidence means 19 times out of 20, this image will be correct”.
- **Embed limitations in a scenario.** The XAI interface could use a percentage to indicate a risk or uncertainty. This percentage could be accompanied by a scenario illustrating weaknesses related to the algorithm, e.g., the high likelihood of a particular bias.
- **Example.** The XAI interface could provide metrics with an example from the past of a similar situation, a similar decision being made, or a similar system being used.
- **Source.** The XAI interface could inform the user about the source of the data or algorithm (this would depend on the domain and the decision). This could be provided as additional material in low-time-pressure situations. For example, an interface could include a link to the database and inform a user about related risks by illustrating the data journey and where the information arrived. This could help experts calibrate their trust in this particular case and see if there are any untrustworthy sources or inconsistencies in the data journey.

Design Goal 4: To slow decision-making down and enable analytical deliberation - to encourage the decision-maker to actively engage with the explanation and support a more thorough analysis

The XAI interface features should help experts engage in analytical thinking and avoid making fast decisions solely based on their intuition, which could lead to overconfidence bias or overreliance on AI. This is particularly important in sensitive cases where experts' intuition cannot be effectively used due to the lack of additional information. To increase engagement, the interface design should cause friction and slow down decision-makers. The XAI interface design should also support experts by providing additional context that would enable in-depth analysis and help experts recognise the similarities and differences of this case so they can compare them to familiar cases. The workshop participants' suggested design features were grouped into four categories.

Add interactive features

Interactive XAI interface features could allow the user to explore and question the data and algorithm and discover variations of outputs and outcomes of different decisions, supporting their analysis of the case. Interactive features could be especially useful in low-time-pressure situations. They could help the user spend more time analysing the output, learning about the model, and understanding the logic behind its working.

- **Manipulating data and making changes.** The XAI interface could allow users to make changes by manipulating the contributing features and observing how these manipulations influence the final output. For example, there could be an editing scale that experts could use to adjust the importance of the factors or suggest new factors or request the information they are missing. Another way could be allowing a user to flag incorrect or incomplete information.
- **Interactive visualisation features.** The XAI interface could provide interactive features, allowing users to interact with complex data visualisations and, in turn, better understand them.

Provide context

Experts could be supported by providing them with more contextual information, in particular, the information that experts identify as necessary for them to be able to make an informed decision. Simplifying the whole decision-making process, including data gathering and data analysis processes, to a brief output limits experts' decision-making. Contextual information could allow experts to connect information and compare output to similar cases. Providing context could allow users to think about the output in a particular scenario, stimulating their expert knowledge and ability to compare the existing situation with past experiences. Contextual information could be particularly useful in low-time-pressure situations when experts have to explore, analyse and ask for more regarding contextual information.

- **Contextualise with images.** The XAI interface could use images to illustrate the context. Images should be relevant, i.e., show similar to a given cases and be domain- and task-specific.
- **Contextualise with examples.** Examples could accompany explanations. Examples should be relevant, i.e., show similar to a given cases and be domain- and task-specific.
- **Contextualise with stories.** Explanations could be enriched using stories or scenarios. These stories and scenarios should be relevant, i.e., portray cases relevant to the case in question.
- **Use an overlay to show more detail.** The XAI interface could have an optional overlay, allowing experts to explore the context briefly. They could remove it when needed and keep it when unsure about something. The overlay could be a mix of images, examples, or stories. It can also be more structured information, such as rule-based instructions or official schema, which should be followed.
- **General overview and additional learning materials.** The XAI interface could use a general overview of the AI process and provide information about the broader implications. This information should be available for users to read at any time, not only in the moment of decision-making.

Collaboration among experts

XAI interface could allow experts to collaborate and ask for a second opinion in unusual cases or cases where they feel unsure, when the decision is particularly risky, or when the uncertainty is very high. This could increase learning potential and decrease the risk of overconfidence bias.

- **Review screen.** The XAI interface design could include a review screen. This screed would allow users to assess the outcome through a review process. For example, a user might want to invite another user to the review process and get a second opinion before making the final decision.
- **Up/down voting mechanism.** The XAI interface design could include an up/down voting mechanism allowing team or community members to express their opinions. This could create a shared ranking of an algorithm, model, or output.
- **Triangular decisioning.** The XAI interface could allow experts to request another person to be included in a decision-making process, creating a triangle of decisioning (i.e., the main expert + AI + a supporting expert). This could be useful when the decision is particularly sensitive.
- **Peer-to-peer trust networks.** The XAI interface could allow users to access peer-to-peer networks. The network of experts would work as an information hub, and the search would connect users with other experts who are most likely to know the answer.
- **Error log.** The XAI interface could use an error log, allowing users to log their mistakes, for example, in training data or a specific catalogue, creating a friction network that would be available to other users.

Input/Action

In situations when expert engagement is crucial, for example, when a certain decision is particularly risky, the XAI interface could require deliberate action from the user. The input would be asked for at various decision-making stages to ensure that experts are engaged with the information provided and are aware of potential risks and biases. This approach could be helpful in both high- and low-time-pressure situations. In high-time-pressure situations, this approach could help to avoid costly

mistakes and slow down the decisioning to apply the necessary amount of evaluation of the output.

- **Checklist.** The XAI interface design could provide a checklist for the user to complete before making a decision. Using checklists could help to assess the output compatibility with a particular case or scenario and find the best match. This approach could facilitate the learning process.
- **Scorecards.** The XAI interface could include scorecards for all the results/data points/outputs. Scorecards would help users compare various aspects, such as accuracy and quality of information.
- **Time delay.** The XAI interface could assign a time delay for checking and verifying the information, so experts would have to slow down their assessment and avoid quick *accept* or *decline* decisions.
- **Verify before proceeding with a decision.** The XAI interface could include a feature asking users to check and confirm their decision before proceeding further or completing the task. For example, the interface could include a pop-up message requiring users to stop, press a button, or type their responses.

5.8 Chapter overview

This chapter identifies several core ways in which AI-driven decision support systems and explainability fail to meet domain experts' cognitive needs and argues for the importance of psychology-driven differentiated explainability support/design for expert users. Drawing from HFE and Cognitive Psychology research literature, it first demonstrates how expert decision-making strategies can be determined by the levels of users' expertise, contextual risk and time-pressure. It outlines the key decision-making strategies and cognitive biases linked to these aspects. It demonstrates why and how introducing intelligent systems can increase cognitive load and disrupt decision-making.

To address these issues, this chapter presents the ERT Framework, a domain-agnostic conceptual framework developed to inform explainability researchers and guide interface designers by providing information about critical decision-making strategies to support any decision-making scenario. The framework highlights three core dynamics that most heavily influence the manner and type of explanation required and makes indicative suggestions for how designers might begin to think

about explainability in these contexts. It offers a practical means of mapping significant contextual factors to appropriate approaches to explanation and, in this way, provides a pragmatic starting point for interface designers to rapidly incorporate explanations that are most likely to meet user needs and improve system understanding.

The framework was applied to a worked example of AI in journalism, where it demonstrated high applicability in practice. The speculated scenarios also helped to identify the ERT Framework's additional benefit. They showed that even in dynamic environments, where risk, time, and expertise might fluctuate and applying the framework might be complicated, using it can still increase the designer's awareness of the user's needs and differences. It can nudge them to think about the decision-making context from a different perspective. This will be of relevance to both explainability researchers and UX practitioners.

Finally, the chapter outlines a number of interface design suggestions brainstormed by experienced UX designers. These suggestions are applicable in high-expertise, high-risk contexts and various time-pressure situations. Overall, this work provides a novel, user-centred approach to explainability in decision-making based on tailoring explanations to facilitate using experts' naturalistic decision-making and sensemaking strategies.

Reviewing the HFE and Cognitive Psychology literature and introducing the ERT Framework is the first step towards understanding and acknowledging the changes that AI introduces in expert decision-making. This work invites XAI researchers and UX designers to recognise that these changes might disrupt decision-making. It also suggests how explainability could dampen or reverse the effects of this disruption.

Chapter 6 Extended value of explainability: Strategies to narrow down the expertise gap and support learning and engagement

Chapter 6 discusses the extended benefits of explainability and outlines strategies that could be employed to encourage experts to engage with explanations meaningfully. The proposed techniques are also intended to inform explainability interface design and explanations so they would support expertise development and continuous learning. Finally, in this chapter, I propose a nascent intermitted explainability method that could help determine when explanations should be cognitively challenging and when they should be seamlessly embedded into the experts' workflows. This chapter is based on the literature reviews referred to in the previous chapters and supported by additional research literature on learning and expertise development from Cognitive Psychology and HFE disciplines.

This chapter explores the aspect of *meaningful engagement* - one of the three properties of the holistic XAI approach proposed in this thesis.

6.1 Chapter introduction

The HFE literature suggests that with increased automation, experts' motivation and level of engagement are reduced (Sheridan, 2012). If decision-making with AI support mainly involves accepting and rejecting outputs, experts can become passive AI supervisors rather than active decision-makers (Lee & Seppelt, 2009). This passivity increases the risk of expert demotivation and aversion to AI systems (Meske et al., 2022). Furthermore, HFE research suggests that disrupting user decision-making and reducing the amount of engagement required from experts can deteriorate skills involved in making decisions and evaluating information (Bainbridge, 1983). This could mean that experts would struggle to apply their expertise effectively with and without AI's support. Moreover, using intelligent systems provides fewer opportunities to develop expert skills than making decisions without AI help (Schemmer et al., 2021). This could disadvantage novice decision-makers. There is a risk that AI would widen the gap between *new* experts who have technical skills but lack domain expertise and *old* experts who might not have technical backgrounds but are experienced in their domain.

Explainability could be designed to address these issues. However, the literature reviewed in Chapter 2 revealed that experts lack the motivation to engage with

additional AI information, such as explanations, when it is introduced into their already busy workloads and are reluctant to learn information outside their domain (Gu et al., 2020). Thus, explanations are often seen as an additional layer of complexity and are rejected as not applicable, especially if they require experts to apply new skills to interpret them (Conejero et al., 2021).

Explanations can also be shallow. They are often used to briefly explain the output or underlying workings of the system but do not intend to develop a deeper understanding of its workings. As a result, experts rarely see explanations as helpful (Bhatt et al., 2020) and do not meaningfully use them (Jacobs et al., 2021; Schemmer et al., 2021), which suggests that explainability could become a formality rather than an effective support tool.

In this chapter, I argue that the value of explainability should be extended beyond just explaining AI outputs. I show how explainability could enhance learning and domain expertise development by applying learning and engagement strategies informed by Cognitive Psychology and HFE research. Explanations could also become a way to provide and receive feedback, allowing users to become active participants in improving AI systems and incorporating them into their workflows. One of the critical issues with the current XAI techniques is that experts do not see them as valuable and lack the motivation to engage with explanatory information. Extending the value of explainability and providing experts with more control could help increase their motivation. Moreover, I argue that friction-creating design techniques could be used to inform the XAI interface design and make explanations more cognitively engaging, reducing risks of biases and improving the effectiveness of human-AI collaboration.

6.1.1 Publications

This chapter is based on the publication for the NordiCHI'23 conference, *"XAI for learning: Narrowing down the digital divide between "new" and "old" experts"* (Simkute et al., 2022). I developed the research idea, conducted the literature review and wrote the paper.

6.2 *Shallow* explainability

Domain experts often view explainability as ineffective in supporting decision-making or informing their tasks (Bhatt et al., 2020). They also do not see XAI techniques

enriching their critical understanding of AI limitations (Wysocki et al., 2023) or motivating them to learn and solve problems (Eiband et al., 2018). As a result, they lack interest and curiosity to attend to explanatory information unless it juxtaposes their initial expectations or expert opinions (Eiband et al., 2018). Superfluous explanations are more likely to be ignored (Bussone et al., 2015; Naiseh et al., 2021) or inspected superficially (Eiband et al., 2018; Schemmer et al., 2021). This aligns with findings from Cognitive Psychology showing that available features with no apparent value for the person will not be focused on (Simonson et al., 1994). Domain experts have also been shown to ignore explanations, which they cannot effectively contextualise to fit their domain. For example, if explanations do not use domain-specific vocabulary (Naiseh, Al-Mansoori, et al., 2021), are too simplistic or contain unnecessary detail (Bussone et al., 2015). According to Cognitive Psychology research, people generally minimise the load on working memory by concentrating only on essential and concrete information and ignoring that which is not (Legrenzi et al., 1993; Sage, 1981).

Explanations that are available but lack meaningful value, or '*shallow explanations*', might also lead to habitual responses. According to the habit heuristic, people prioritise familiar alternatives that they have previously accepted for a similar purpose (Sage, 1981). People tend to repeat the sequence of actions and choose the same alternative when confronted with a similar explanation or problem (Sage, 1981). Thus, experts might develop a habitual response (e.g., skip or accept) to explanations if they encounter them as a formality rather than a source of valuable information.

Overall, the standard XAI often only provides an *on the spot* short-term solution but loses its initial value in the long term (Bansal et al., 2021; Lee et al., 2017). This increases the risk of XAI becoming a formality. If explanations are seen as an additional redundant feature, they are more likely to be ignored or inspected superficially (Eiband et al., 2018; Schemmer et al., 2021).

6.3 Technical skills gap between *new* and *old* experts

AI-driven decision support systems are used in various domains, where different decision-makers represent varying technical skills (e.g., expertise in data analysis, AI, mathematics, statistics, or data visualisation) and domain knowledge. Here, technical proficiency in decision-making is defined as *new* expertise and domain-

related knowledge and skills are defined as *old* expertise. Poor technical skills might prevent *old* experts from uncovering valuable information and effectively benefiting from the available technologies (Gilvary et al., 2019). Moreover, without effectively understanding the workings of AI, domain experts are more likely to succumb to confirmation bias and trust the model when its predictions align with their expert knowledge (Bayer et al., 2021).

In some areas, AI literacy might be a prerequisite, but research by journalists (Jones & Luger, 2021) and public sector workers (Brown et al., 2019) has shown that domain experts often lack even a basic understanding of AI. The level of AI knowledge among experts can also vary within the same domain. For example, journalists working for small news agencies or local branches of big news agencies have less training and exposure to AI systems (Zhang et al., 2020). Variation in technical knowledge can also result from changing prerequisites to enter job positions and even changing university curriculum for some professions (Renz & Hilbig, 2020). A high level of one skill, but not the other, might result in disrupted trust dynamics. Higher domain expertise has been shown to negatively affect trust in AI-driven decision support systems (Bayer et al., 2021). Similarly, higher levels of AI knowledge have been shown to result in an algorithmic aversion (Jacobs et al., 2021). On the other hand, a lower level of technical skills can result in a sense of illusory superiority and rejection of advice or explanation despite needing it (Schaffer et al., 2019).

6.4 Risk of deskilling in AI-supported decision-making

Experts' ability to intuitively grasp situations and perform in a fluid, flexible, and highly proficient way are unique and valuable assets for decision-making (Klein, 2003). Expertise can increase the quality of decisions, as the latter depends on how the decision-maker can acquire and analyse provided information and how well they can evaluate and interpret it to discriminate between relevant and irrelevant aspects (Sage, 1981). Experts can be distinguished from novices by their ability to effortlessly perceive information beyond its surface features and recognise meaningful patterns and exceptions to typical situations with high proficiency (Dreyfus & Dreyfus, 1986). Experts can also step out of the intuitive decision-making role and make more deliberate decisions (Hutton & Klein, 1999).

One of the most critical aspects of expertise development is the experience or time spent performing a task (Hutton & Klein, 1999). When AI is introduced in a decision-making context, experts do not have the same opportunity to practice their decision-making skills. This means they rely on AI recommendations but do not have to analyse, gather and process information themselves (Buçinca et al., 2021). This change in decision-making can negatively affect expertise development and even lead to the deterioration of skills (Bainbridge, 1983). From a long-term perspective, this can mean that valuable human input and unique expert skills that AI cannot replace, such as intuition and pattern recognition, would be potentially lost (Bainbridge, 1983; Klein et al., 2007).

The implementation of AI systems often means that experts have to monitor AI's performance, for example, notice and override erroneous recommendations (Green & Chen, 2019a). Passively watching processes without actively applying expert skills stifles experts' abilities to retrieve knowledge from long-term memory, which depends on the frequency of use (Bainbridge, 1983). This might also prevent experts and novices from generating new strategies for dealing with unexpected situations, especially if the system is opaque and does not provide sufficient contextual information and feedback about its performance (Bainbridge, 1983). Furthermore, this could lead to the decision-makers' aversion towards AI-supported systems (Meske et al., 2022). HFE research shows that task automation demoralises highly skilled workers and decreases their motivation (Lee & Seppelt, 2009). When experts are resigned to the fact that they cannot add value to automation guidance, they lose motivation and are less likely to learn new skills which they could apply in future decisions (Stuart et al., 2007). Skilled practitioners might feel like novices again due to the loss of skill or because the system is unfamiliar and requires an entirely new skill set (Bainbridge, 1983). If experts habitually overrely on AI predictions, the persistence of automation bias might also result in reduced cognitive engagement and the development of faulty mental models (Glikson & Woolley, 2020). This threatens to deskill experts, especially in highly automated tasks (Schemmer et al., 2021). AI-related deskilling risk emphasises the importance of supporting experts' continuous learning and expertise development.

6.5 Effective human-AI collaboration

AI systems have the potential to enhance experts' work. For example, AI can provide information supplementing experts' knowledge, which can help them to recognise unique cases and out-of-ordinary patterns across large volumes of data (Bolander, 2019). AI systems can also help experts integrate knowledge from different sources, make initial insights and impressions of the situation, and intuitively link those to the potential outcomes (Procter et al., 2023). In addition, AI systems can support experts' ability to deal with uncertainty and incomplete information (Hong & Lee, 2018; Wang & Yin, 2021; Zanzotto, 2019) and think beyond their existing knowledge, leading to more thoughtful and creative decisioning (Tan et al., 2018). AI can also provide a more systematic approach to decision-making. It can help decrease the risk of human decision-making biases and reduce the variability across decisions (Sousa et al., 2019; Zhang et al., 2020). AI systems can be used to automate mundane tasks (Janssen & Kuk, 2016) or tasks where AI outperforms human capabilities (rule-based and repetitive tasks), allowing experts to focus on more sensitive or creative parts of their work (Haenssle et al., 2018).

However, experts often perform worse when working with the support of AI systems than if they were performing the same tasks alone (Sheridan, 2012). HFE research literature suggests that when automation is introduced in experts' workflows, they do not shift their attention effectively (Altmann & Trafton, 2002; Cork et al., 1998; Janssen et al., 2015). Experts can experience increased cognitive load and feel overwhelmed by the new *role* of supervising opaque AI systems (Baxter et al., 2012; Sheridan, 2012). Thus, instead of benefiting from the provided information, they are investing their cognitive efforts and time into monitoring automation performance for errors and evaluating the trustworthiness of its outputs (Grubb et al., 1995; Warm et al., 2008). On the other hand, if they perceive intelligent systems as superior to or as reliable, they are more likely to lose vigilance and be distracted than attentive (Bainbridge, 1983; Jones & Endsley, 1996). HFE findings suggest that experts need support to apply the additional information provided by the intelligent systems and effectively use freed-up time and attentional resources (Lindgren, 2023; Metzger & Parasuraman, 2005). The effective collaboration between domain experts and AI should be supported by building an environment where their performances complement each other. Experts should be able to overturn and fine-tune the results

of AI (van Baalen et al., 2021). Engaging explainability could be used to enable experts to make effective use of AI and create this type of environment.

6.6 Extended value of explainability

The value of explainability could be extended beyond just explaining AI outputs or the underlying workings of the system. Using design strategies that engage and motivate learning could help experts benefit from intelligent systems. Several studies already demonstrated how XAI can help users learn about AI workings (Mirchi et al., 2020; Schneeberger et al., 2019) and improve their awareness of potential biases (DeVos et al., 2022a). Sheh (2017) proposed that AI's knowledge gathered during training can also be used to educate its users creatively (Sheh, 2017). These and similar examples in this section illustrate that explainability could be designed to support more than just interpretations of AI outputs and could have an extended and lasting educative effect.

6.6.1 Learning via XAI feedback

Research suggests that explanations providing feedback regarding the reasons for the success or failure of the system can significantly help users learn about it. Especially if the learner can interrupt the AI processes, ask questions, and request feedback (Fiok et al., 2022). Allowing users to interact with the system and explore various process steps, such as visual metaphors, can help them understand how it works (Misra et al., 2021). AI has already been shown to recognise different aspects of expert surgeon performance and use this knowledge to provide feedback to junior surgeons. This feedback could help train new surgeons using explainable tools such as virtual operative assistants (Mirchi et al., 2020). XAI techniques have also been shown to help users provide feedback that would allow the integration of their domain knowledge into the systems (Misra et al., 2021). Interacting with the system through explanations and feedback has been shown to improve pathologists' ability to engage with the outputs of a model actively and, in turn, provide feature-based feedback that can be used to refine it (Hun Lee et al., 2023). Hun Lee and colleagues (2023) showed that XAI can create collective hybrid intelligence on a complex decision-making task with improved accuracy and consistency. Feedback via interactive XAI can also be influential in sensitive domains, such as computational psychiatry, where affected users often miss meaning and relevance in

the AI predictions (Roessner et al., 2021). AI systems, followed by adequate explanations, have also been shown to help users train safely and control difficult social situations or high-risk tasks (Schneeberger et al., 2019).

6.6.2 Increasing awareness of biases

Explainability can improve fairness judgements and even educate users about various biases. DeVos et al. (2022b) demonstrated that users can detect and audit potentially harmful algorithmic behaviours when supported with explanations. The authors argued that XAI could be used to introduce user-driven algorithm audits and help leverage users' prior experience of harms and biases (DeVos et al., 2022a). XAI also has shown the potential to educate young users. Hitron et al. (2019) indicated that children could learn technical concepts by experimenting with an explainable model and notion of bias. The ability to interact with an explainable model significantly improved 10-to 14-year-olds' understanding of the concept of bias in terms of gender discrimination (Pérez & Isaac, 2020). When XAI enabled decision-makers to make holistic risk assessments and adjust for algorithmic limitations, workers could reduce the racial disparity in child welfare screening compared to AI, which was used without explanations (Cheng et al., 2022). Moreover, effective XAI techniques have been shown to help users reflect on their own decisions and potentially reduce the potential for biased decision-making (Bansal et al., 2021; Roessner et al., 2021).

6.7 XAI for learning and domain expertise development

Explainability could be designed with the intention to increase the potential for explanations to provide lasting knowledge development. HFE and Cognitive Psychology research literature is rich in strategies that could inform the design of more engaging and motivating explanations. Based on this literature, a list of strategies was synthesised and aligned with AI contexts. These strategies are intended i) to improve the educative potential of XAI and ii) to improve a meaningful engagement with explanations. Research shows that effective explanations need to be cognitively engaging and valuable, meaning decision-makers should be nudged to attend to them consciously (Buçinca et al., 2021). Explanations should also provide relevant and helpful information to maintain users' enthusiasm and motivation (Goddard et al., 2014). Combining engagement and learning methods

could expand the educational potential of XAI and support domain knowledge development, preventing AI-related expert deskilling. This could be especially beneficial in less technical domains, where experts might lack the computational knowledge to interpret AI outputs and explanations (Chouldechova et al., 2018; Woodruff et al., 2020).

6.7.1 Cognitively engaging versus shallow explanations

The reasoning behind cognitively engaging explanations is that they should help transfer new information from short-term to long-term memory, supporting lasting learning. Short-term memory is essential for an immediate recall, but without a conscious effort to recall the information, it is argued to be lost in less than a minute (Shiffrin & Schneider, 1977). To transfer the information from short to long-term memory, constant conscious attention via rehearsal is needed (Sage, 1981). This process is a base for information processing. Building lasting knowledge that could be later retrieved and used in future decisions requires attention to the aspects of the incoming data (Sage, 1981). Shallow explanations do not require users to pay attention, so they are unlikely to study explanations intentionally, trying to remember and later retrieve provided information, especially if they perceive XAI as a redundant feature. Individuals tend to use the least cognitive effort when supported by automation, especially if a large part of the task is automated (Parasuraman et al., 2000). They are more likely to rely on cognitive shortcuts and only superficially attend to the provided information instead of using analytical thinking (Evans & Stanovich, 2013). Thus, *shallow* explanations might not have lasting educational value and support information transfer from short-term to long-term memory. Moreover, passively monitoring the performance of AI can feel demotivating (Goddard et al., 2014). Providing experts with more agency via engaging explanations (e.g., allowing them to give feedback or feel a sense of growth in learning) could lead to increased motivation to thoroughly evaluate explanations.

6.7.2 Reflecting on expert knowledge

One way to make explanations more engaging is to increase their relevance to experts. Explainability should fit with the experts' workflow and use domain-specific terminology (Anjomshoae et al., 2019; Dikmen & Burns, 2022). Research shows that the lack of relevance is one of the reasons why experts ignore explanations

(Bussone et al., 2015; Naiseh et al., 2021). Experts are more likely to use explanations when they reflect their information needs in various contexts (de Greeff et al., 2021; Lukyanenko et al., 2021; Schaekermann et al., 2020). For example, Tonekaboni et al. (2019) study with clinicians using explainable AI 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 (Tonekaboni et al., 2019). Seeing that allowed experts to compare AI outputs to their clinical judgement, especially in case of a disagreement (Tonekaboni et al., 2019).

Experts find it easier to reason when AI-provided information aligns with their existing knowledge than if it requires learning new concepts (Gu et al., 2020). For example, explanations that provide supplementary material that aligns with the information experts seek (e.g., the patient's medical history or similar cases) could help them make informed decisions and benefit from AI support (Gil et al., 2019). Using design features that prompt users to reflect on their prior knowledge could foster engagement and consolidation of their expert knowledge and ability to apply it when making decisions in collaboration with the AI (Dudai et al., 2015). By reflecting on their specialist knowledge, users could also gain consistent experience in domain-specific decision-making (Klein, 2008). Linking terms representing the contributing features to the domain-relevant context and customising explanations to the needs and requirements of experts have been shown to be effective in supporting mental health practitioners (Yang et al., 2021; Zhang et al., 2020) and physicians (Naiseh et al., 2021).

6.7.3 Cognitive forcing

One of the most popular techniques used to increase user engagement is cognitive forcing. Cognitive forcing is a technique forcing or nudging a user to stop before they make the final decision and to perform an additional task, for example, to conduct a secondary search (Croskerry, 2003) or explain the reasoning behind their decision and justify why they did not choose an alternative (Buçinca et al., 2021). It promotes analytical thinking and self-monitoring by disrupting heuristic reasoning (Lambe et al., 2016). Forcing users to slow down and mindfully attend to the provided information can prevent fast and superficial evaluations. The *fast* heuristic-based decisioning has been shown to lead to an overreliance (de Visser et al., 2020).

Cognitive forcing is intended to challenge decision-makers' heuristic reasoning and make them think in a *slow,* analytical way (Kahneman, 2006). Using cognitive forcing techniques, requiring decision-makers to attend to provided information, could decrease the risk of cognitive biases, such as confirmation bias (Sage, 1981), and help to mitigate overconfidence (Schaffer et al., 2019). In the context of XAI, explanations that challenge and require this type of engagement can be more effective in reducing overreliance than regular XAI approaches, but they can also frustrate experts. For example, Buçinca et al. (2021) used three cognitive forcing options in their study. Participants had to either deliberately select an option to see an explanation, update their decision after receiving it or experience a delay before seeing it. Participants were more attentive to cognitively engaging explanations than to the regular ones, but they disliked their design (Buçinca et al., 2021). Thus, the value of engaging with explanations should be clearly communicated to the experts to avoid their premature rejection.

6.7.4 Explanations providing context

Using additional (domain- and task-specific) context to enrich explanations could increase their applicability and value. For example, providing information about the importance of various features could make them more meaningful and comprehensible to specific experts (Bove et al., 2022). Finding what information is meaningful to experts might require involving them in the design process. For example, ask them what aspects they consider when making decisions and how they evaluate them. Adding domain-relevant information to the explanations could make them more relevant and, thus, more motivating, which is one of the criteria for experts to be willing to attend to explanations (Naiseh et al., 2021). In the healthcare sector, explanations that explain the output and provide medically important information could foster experts' ability to recognise similar medical cases in the future (Gu et al., 2020). Adding relevant contextual information could help experts relate to explanations better (Gil et al., 2019) and help them apply their past knowledge to the current situation (Dudai et al., 2015). Including domain-relevant information to contextualise explanations could improve experts' satisfaction and overall performance (Bove et al., 2022). Providing additional context through explanations could also help experts interpret AI outputs more effectively and improve their understanding of the system's behaviour (Baudisch et al., 2002).

6.7.5 Explanations using analogies and scenarios

Explanations could be made more engaging by using analogies and mock scenarios to support a mental simulation of potential decision outcomes. Cognitive Psychology research suggests that analogies can help develop and utilise appropriate information evaluation heuristics (Sage, 1981). Using analogous comparisons to define alternatives to the prediction could also stimulate expertise development. HFE research has shown that expertise development is stimulated by evaluating various outcomes, comparing them, and using reasoning abiliites (Klein, 2008). Analogies and scenarios could also enable experts to recognise similarities to previously experienced situations, enabling the practice of existing expert skills (Dreyfus & Dreyfus, 1986). Explanations providing examples of potential outcomes are more likely to be evaluated using mental stimulation than the ones providing a comparison of a generic set of criteria (Hutton & Klein, 1999). Thus, they could increase analytical evaluation and reduce the risk of cognitive biases.

The use of scenarios could also increase the quality of decisions. Using scenarios of potential outcomes paired with sufficient instructions has been shown to increase decision accuracy (Koehler & Harvey, 2008). Scenario-based explanations could provide decision-makers with information needed to enhance their perceptual skills, enrich their mental models about the domain, construct an extensive and varied repertoire of patterns, and provide a more extensive set of routines based on past instances (Koehler & Harvey, 2008). Scenario-based explanations could not only allow experts to practice their expert skills continuously but also introduce novices to case examples (Koehler & Harvey, 2008) and expert-like decision-making (Hoffman et al., 1998). This could increase novices' exposure to the situations necessary for their expertise development.

6.7.6 Explanations providing contrastive concepts

Explainability educational potential could be increased by applying contrastive concepts. This method is often seen as an effective method for training medical students. For example, teaching diagnostic categories by contrasting them, rather than asking students to learn individual diagnoses, has been shown to help them associate each disease with its discriminant features (Klayman & Brown, 1993). Students who learned by using this method also made better diagnostic judgements

than the students who learned about each disease individually (Klayman & Brown, 1993). Contrasting information and juxtaposing features rather than showing each feature independently has been shown to teach individuals to be vigilant to features that have particular importance and could inform their future decision-making (Klayman & Brown, 1993). This method could help to highlight distinctive output features and develop the ability to notice even salient out-of-ordinary events that should be further investigated (Hutton & Klein, 1999). Emphasising aspects in which characteristics differ and demonstrating that there might be alternative reasons for predictions can help to develop their conceptual thinking (Johnson-Laird & Shafir, 1993). It can also help distinguish typical features from those that are particularly important (Klayman & Brown, 1993), avoid habitual responses to predictions (Sage, 1981), and use recognitional decision-making strategies (Klein, 2015b). Contrastive concepts have also been shown to be easier to retrieve (Johnson-Laird & Shafir, 1993).

6.7.7 Explanations enabling feedback

Allowing experts to have more agency in AI-supported decision-making has been shown to reduce algorithmic aversion (Dietvorst, 2016) and increase user motivation (Naiseh et al., 2021). One of the ways to give users agency is to allow them to provide feedback. Explainability interface design should include an affordance allowing experts to give feedback directly. For example, if they override a recommendation, they could give feedback on why it was done and how the AI system could be improved. They must be informed about the changes that their feedback has influenced to increase their motivation to continue providing it. Interacting with the system through explanations and feedback has been shown to improve pathologists' ability to engage with the outputs of a model actively and, in turn, provide feature-based feedback that can be used to refine it (Hun Lee et al., 2023; Roessner et al., 2021). Interacting with explainability through feedback could also create collective, hybrid intelligence on a complex decision-making task with improved accuracy and consistency (Hun Lee et al., 2023). Moreover, it could encourage experts to engage with explanations mindfully and promote effortful reflection and a more thorough analysis of the AI outputs (Ehsan & Riedl, 2020).

6.7.7.1 Feedback to the users about their performance

Explanations could include feedback to the users. For example, explanations could incorporate feedback informing users about their performance or how their actions influenced AI's behaviour (Battaglia et al., 2003). Establishing a regiment of feedback has been shown to facilitate expertise development (Battaglia et al., 2003), as simply practising a task without receiving feedback is usually insufficient for expertise building (Salas et al., 2002). Receiving feedback allows experts to monitor their progress, strengthen their intuition (Battaglia et al., 2003) and improve their decision-making skills (Klein et al., 2006a). Explanations with feedback can improve decision-makers' self-awareness (Naiseh et al., 2021) and prevent overconfidence (Klein et al., 2006a). Explainability that gives feedback to users regarding their performance could help experts reflect on their own decisions and reduce the potential for biases (Bansal et al., 2021). Incorporating examples of how decision-making could be improved into feedback could lead to even more significant learning potential (Schneider, 1985).

6.7.7.2 Feedback to the users about AI performance

Explainability could also be designed to include feedback about the system's performance. For example, it could inform users about contributing features and their weights towards the output. The feedback that emphasises relationships between influencing factors has been shown to improve expert judgements in human-automation situations (Balzer et al., 1989). Feedback could build a better understanding of contributing features and improve expert judgements (Balzer et al., 1989). Learning from feedback is particularly effective when users are able to interact with an automated system instead of just passively observing it (Hoffman et al., 1981; Klayman, 1988; Klayman & Brown, 1993). For example, if experts can choose the combinations of cues and instances of information to test instead of just seeing the representative instances (Klayman, 1988). Seeing task characteristics, such as feature values and response characteristics, individual cue values, and associated subject responses, is an effective way to learn about complex tasks and their relationships (Hoffman et al., 1981).

6.7.8 Interactive explanations

Simulating AI-generated outputs can aid expertise development and help to build a better understanding of the system's workings (Klein, 2008). Interactive explainability features can enable a mental simulation of various decision options, help to develop new ones, allow to more effectively predict their outcomes and assess impacts (Klein, 2008). Being able to manipulate data points and interactively explore explanations can provide opportunities to practice expert skills and, in turn, lead to the formation of more complete mental models about the domain (Smith et al., 2004).

Interactive features could also support learning about the technical aspects of the system. For example, manipulating the model behaviour by adding class labels and tuning the classifier's parameters can improve system understanding (Höferlin et al., 2012). Interactive visualisations (e.g., manipulation of visual elements and data items) can be particularly effective in aiding users' understanding of the AI model structures (Sacha et al., 2017). Interactively experimenting with visual features can support *what-if* analysis and help identify which data items are critical and how they relate to specific features and the final output (Sacha et al., 2017). Interactive features can also aid experts' ability to experiment with different scenarios of the model outcome and allow a more cognitively engaging, deeper analysis of it (Bohanec et al., 2017). It can also increase the sense of agency and motivation. Making even minimal alterations has been shown to give users a feeling of control and increase their willingness to rely on the system (Dietvorst, 2016).

6.7.9 Intermittent applications of cognitively engaging explanations

The engaging XAI approach requires users to break their workflow and attend to the explanations. This notion challenges conventional design principles of the seamless interface design (Norman, 1999). Moreover, despite improved decision-making accuracy, users rate cognitively engaging explainability as the least enjoyable to use compared to explanations that do not use features such as cognitive forcing (Buçinca et al., 2021). Moreover, engaging XAI requires users to spend additional time deliberating about the explanatory information, learning new information, providing reasoning for their decisions, or generating feedback. Research shows that users reject AI systems and their features that add to their already intensive

workloads (Gu et al., 2020). Thus, these types of explanations might not be suitable in every situation.

I propose an intermittent approach to XAI, which determines cases where these types of explanations should be used. This method is intended to help regulate when cognitively engaging explanations could be helpful and when they could be distracting or even harmful. I also suggest a scoring approach based on contextual factors discussed in the previous chapters and a dynamic adjustment method based on the level and type of expert domain and task knowledge. Finally, I argue, that the value of engaging explanations should be clearly communicated to the expert decision-makers.

6.7.10 Contextual factors

The previous chapter suggests that contextual factors, such as contextual risk, time-pressure, and level of expertise, influence decision-making strategies. These factors could guide the use of engaging explainability. Woodruff et al. (2020) showed that user preferences for explainability change across scenarios, and in some cases, it could compromise decision-making. For example, when time-pressure is high, it could be counterproductive to provide too many information recources and overburden users with additional knowledge that might not be useful for a particular decision, even when it could support their learning over time (Hutton & Klein, 1999). It could be frustrating and distracting because attending to explanations would add to experts' high cognitive workload (Haldane & May, 2011; Oliva & Sterman, 2001). High-time-pressure is already cognitively demanding, and adding extra pressure might result in explanation rejection (Schemmer et al., 2021). The goal in these situations should be to reduce decision-makers' cognitive load. On the other hand, in high-risk situations that are not time-pressured, explanations should be designed to slow down intuitive expert reasoning to avoid biased decisions (Klein, 2008). In these situations, encouraging users to engage with explanations and challenging their heuristic thinking could be necessary, even if they do not enjoy it (Buçinca et al., 2021).

It is also important to consider situations that could be valuable learning opportunities. For example, novel decision-making situations could be rich in information that could facilitate learning for both novices and experienced decision-makers, especially in low-time-pressure and high-risk contexts. In these situations,

experts could invest time familiarising themselves with contextual information and interacting with mock scenarios, analogies, and visualisations. These are also opportunities to provide feedback about the system's performance and design. Similarly, learning opportunities could be found in low-time-pressure and low-risk contexts when situations are unique, and the knowledge gained from interacting with them could be helpful in similar future circumstances. Engaging explainability could feel redundant in low-risk decision-making situations that do not require high expertise or are straightforward. Thus, it should be optional, but users should still have the option to ask for an informative explanation for learning, feedback, and verification purposes.

Based on the HFE literature reviewed in Chapters 3 and 5, I propose categories of decision-making situations that could determine the suitability of engaging explanations.

Engaging explanations should be used:

- In decision-making situations that require analytical thinking and challenging heuristic-driven decisioning, e.g., due to the high-stakes
- In novel decision-making situations, e.g., when the new system is introduced to experts or when the task is unfamiliar
- In decision-making situations that are particularly rare, unusual, or informative, e.g., when decision-makers must supplement their expert knowledge with additional information provided by the AI systems
- When novice decision-makers proactively seek learning opportunities

Engaging explanations should not be used in situations:

- When the cognitive load is high due to high-time-pressure
- When the decision is routine or when the task is particularly easy, and its completion does not require a high level of expertise.

6.7.11 Adapting explainability design

Given that explanations are used to promote learning, they should be dynamic and adapt to the changing level of model understanding. In the case of novice decision-makers, the design of explainability interface should also reflect the development of domain expertise. Research shows that decision-makers find explanations redundant, when they are too simple or repeat information that they already know

(Bussone et al., 2015). Thus, the level of engagement required by the explainability design should gradually reduce with the increased expert knowledge.

The lower level of domain knowledge would require more frequent use of cognitively engaging domain-related explanations. However, with gradual improvement using these explanations and after demonstrating the ability to reason based on the information provided, engagement requirements should be reduced. Schoonderwoerd (2021) showed that decision-makers expect that their understanding of the system will increase over time, and they will receive less extensive explanations. If this dynamic adaption does not happen, the explanations might become frustrating (Gu et al., 2021). Experts need explanations that match their technical backgrounds and information preferences (Schoonderwoerd et al., 2021). If explanations are overwhelmingly detailed, they might lead to automation bias (de Visser et al., 2020). Depending on the level of engagement and performance of the AI and humans, the system could assign a score of understanding and recommend the right level of cognitive friction. A similar approach, called adaptable allocation, has been suggested by Zerilli and colleagues (2022), where decision-makers were kept vigilant by machines assigning tasks of varying difficulty to human decision-makers.

6.8 Chapter overview

Even carefully tailored explainability will not be useful if experts do not attend to it. This chapter presents a novel approach to explainability that would extend its value. Research shows that users are unlikely to learn from explanations that do not require their engagement (Eiband et al., 2018). They are also unlikely to spend time understanding explanations if they are not motivated or do not see value in them (Gaube et al., 2021; Naiseh, Al-Mansoori, et al., 2021). The *shallow* applications of explainability introduce a risk that experts will ignore or inspect the explanations superficially, which could result in automation bias (de Visser et al., 2020). Explanations are rarely designed to promote long-term learning and are often only useful at the moment they are provided. These types of *shallow* explanations could easily become habitual and meaningless (Sage, 1981). The potential for explanations to enhance experts' knowledge and skills is also rarely exercised.

I argue that explanations should also aim to support learning and expertise development instead of only being used for a single output or to rationalise the

system’s inner workings. As discussed in previous chapters, when AI systems are introduced into experts’ workflows, they often cause challenges related to changing decision-making and reduced contextual information. There is also a risk that experts’ skills will deteriorate if they have no meaningful agency in AI-supported decision-making. In addition, novices might have fewer opportunities to develop expert skills. This could widen the knowledge gap between experts and novices. I argue that explainability could be used to address these issues and provide additional benefits. Several studies have already demonstrated that well-designed explainability could positively impact learning about the system and fairness-related risks. In this chapter, I outlined reasons why explanations should be engaging and suggested how they could promote long-term learning and expertise development. In this chapter, I propose that explainability could be designed to be engaging, motivating, and educative by applying the following strategies:

- Reflecting on expert knowledge in explanations
- Using cognitive forcing methods
- Supporting explanations with relevant context
- Using analogies and scenarios
- Providing contrastive concepts
- Enabling feedback
- Adding interactive features.

Furthermore, I propose that engaging XAI design should be used in moderation, depending on the contextual factors impacting users’ cognitive load and learning needs. The required engagement level should also adapt to changing experts' needs and learning progress.

Chapter 7 Discussion and conclusions

Research reported in this thesis demonstrated the complex nature of explainability in expert contexts and the lack of appreciation for these complexities in the XAI research. The literature review of explainability methods, concepts and techniques (Chapter 2) showed a predominantly technical focus of the XAI research field. A few key publications acknowledged the various XAI stakeholders and their different explainability needs. Still, most of the proposed XAI solutions were designed to support users with high AI literacy and data science knowledge (Adadi & Berrada, 2018, Gunning & Aha, 2019, Guidotti et al., 2018) . A few studies with experts showed that they did not view explainability as designed for them and found explanations as either too simplistic or overly complex (Bussone et al., 2015; Bhatt et al., 2020). The literature review suggested that the key XAI methods (e.g., complex visualisations) required experts to have reasonable data science knowledge to be accurately interpreted. For users without this expertise, explanations could become an extra layer for errors rather than a helpful feature. Technical XAI approaches also failed to consider that explainability should empower domain experts and help them to continue providing value to their expert tasks.

Most importantly, the review showed that the XAI literature did not explore challenges related to explainability in expert contexts or address experts' specific information needs. The XAI community also overlooked the dynamic environments in which experts work and how important it is that the new tools fit these environments, especially in high-risk and time-pressure contexts. For example, in the healthcare sector, experts have carefully structured workflows and work patterns, and automation-caused disruptions to these patterns have been shown to result in costly errors (Klein et al., 2006b, 2006a).

The subsequent literature review in Chapter 3 emphasised the importance and limitations of explainability in ensuring algorithmic fairness, transparency and accountability. Experts play an essential role in overseeing AI outputs, and without being able to understand them, they are unable to do it effectively (Amann et al., 2022; Bolander, 2019; Procter et al., 2023). Besides, research shows that experts struggle to exercise their expertise when using AI systems and often perform worse with AI support than without it (Gaube et al., 2021; Micocci et al., 2021). Moreover, when AI or automation is introduced into experts' workflows, it might lead to errors in

their decision-making (Elwyn et al., 2013; Klein et al., 2006a). It can also leave experts with limited information, reduced situational awareness and increased cognitive workload (Lee & Seppelt, 2009; Sheridan, 2012). The review also revealed that without correctly understanding AI outputs, experts cannot build meaningful trust and, as a result, succumb to various cognitive biases. This can lead to automation bias (Gaube et al., 2021; Skitka et al., 1999) or algorithmic aversion (Dietvorst et al., 2015). HFE research further suggests that introducing automation can lead to skill deterioration (Bainbridge, 1983). For example, if experts are left to supervise AI systems without opportunities to apply their expert skills in AI-supported decision-making, these skills might degrade over time. Many of these challenges are currently not addressed by XAI research communities (Conejero et al., 2021; Bertrand et al., 2022). Indeed, some research suggests that explainability often makes them worse (Bertrand et al., 2022; Kaur et al., 2020; Bansal et al., 2021; Buçinca et al., 2021).

In Chapter 3, I demonstrated the importance of understanding what difficulties experts experience when interacting with AI systems. I developed a knowledge base on AI-expert trust dynamics and reported the essential challenges of the human-in-the-loop approach. I also provided an overview of XAI opportunities and limitations in addressing these challenges.

In the subsequent chapters, I explored how explainability could support the complex needs of experts. I demonstrated that this topic should be approached from a broader perspective and meet several fundamental requirements to be effective. I defined this approach as *holistic* because it encompasses multiple parts, not just an explanation. Here, I argue that the fundamental steps of AI introduction should be covered before introducing XAI features to experts. First, I showed that experts should be able to use explainable AI systems, and as the research findings in Chapter 4 suggest, they currently are not. Research conducted at the large science centre revealed that experts often struggle with practical hurdles that are not seen as necessary by other stakeholders, such as software engineers. The study showed that seemingly simple frictions, such as not knowing how to install the software or what command line to input, prevented experts from using these systems altogether. These frictions led to them being unable to use AI technologies to support their work. At the same time, software engineers made assumptions about why AI adoption is low, sometimes seeing a lack of explainability as the main issue. However, the research findings revealed that AI adoption challenges might stem from a lack of

support, training, and collaboration between stakeholders. Study participants explained that they spent most of their time trying to understand how to use the system, and when asked about explainability potential, they said they had not had a chance to think about it. This shows that it is essential to provide support and training to ease experts' journey into adopting new AI tools. By gradually building foundational solid knowledge, experts would be better equipped to understand explainability solutions later in the process.

The findings reported in Chapter 4 suggested that practical hurdles could be resolved if experts initially received support when using a new AI system. I suggested that this support should be introduced as an introductory period. This way, they would learn by practice and could establish troubleshooting skills and the confidence to ask questions. Moreover, I proposed that experts should receive training on using a new AI system, which should be developed and updated in collaboration with them. Finally, they should be made aware of the capabilities and limitations of AI systems, and this information should be communicated to them in a way that is relevant and understandable. Overall, the study findings in Chapter 4 showed the importance of collaborating with experts when developing and introducing tools designed to support them. In this study, collaboration between software engineers and science experts led to experts being more involved, motivated and proactive when exploring new technologies and troubleshooting them. Software engineers became more aware of expert needs. They adapted their communication style to explain aspects of their developed technologies in a way relevant to the experts.

In Chapter 5, I demonstrated the importance of another fundamental aspect of the holistic explainability approach. I showed that explainability should be designed to enable experts and help them exercise their expertise. Based on the HFE literature, I showed that the explainability interface design should be tailored to support experts' sensemaking and decision-making strategies and reflect contextual aspects, such as time-pressure and risk. When reviewing HFE research literature, it became apparent that experts perform well when they can flexibly make decisions, exploring the available information in their preferred ways (Klein et al., 2010). They can also best apply their expert skills when they have access to information that would help them contextualise the automation outputs and trigger familiar cues and patterns (Klein, 2003; Klein et al., 2010). However, in high-risk contexts and when

working under time-pressure, experts can overrely on their intuition and make biased decisions (Langer, 1975, Svenson, 1979). In these situations, they perform better if given some structure and rules, and their decision-making is slowed (Svenson, 1979; Orasanu, 2010; Dobrow et al., 2006). Explainability solutions are not tailored to accommodate these traits. In response to this knowledge gap, I developed a conceptual ERT Framework that could be used as guidelines to inform XAI interface design in various contexts. I populated this framework with design solutions that the user interface designers had proposed during the ideation workshops. These guidelines are intended to provide actionable design solutions that could inform XAI designers and researchers and increase the usability of existing explainability techniques. Moreover, this chapter reports on the two speculated journalism case scenarios in which the ERT Framework was embedded. These scenarios revealed that my proposed framework could encourage designers to reflect on users from a broader perspective by also considering their work context and their level of expertise.

Chapter 6 addressed another knowledge gap I identified while reviewing XAI and HFE literature. As argued in Chapter 3, introducing new technologies can interfere with users' ability to develop and retain expert skills, as it strips the decision-making process of contextual information and reduces the role of experts in monitoring AI outputs. However, explainability often only provides short-term solutions by explaining an instance or an output but does not increase users' AI literacy or domain-related knowledge. Experts also often ignore or do not attend to explanations appropriately. Research suggests that domain experts do not see a clear value in explanations and instead view them as an additional task or an extra layer of complexity. In this chapter, I proposed that explanations should be designed to help transfer new information into long-term memory, extending explainability's actual and perceived value. Reflecting on the learning strategies from the Cognitive Psychology research literature, I provided a list of design solutions that could use explainability to increase expert knowledge. I suggested that the XAI interface design should include features such as various analogies or additional domain-relevant information. I also outlined strategies to make explainability more engaging and ensure that experts attend to them, for example, by using cognitive forcing techniques and creating design friction. I argue that encouraging experts to engage with explanations would make them think about their decisions more deliberately.

They would also more thoroughly evaluate the available information and the technology they use, which could help tackle the issue of expert deskilling and increase experts' ability to notice errors.

7.1 Contributions of the thesis

This thesis makes numerous novel contributions. First, the most considerable contribution is the new holistic approach to explainability in expert contexts. In this thesis, I proposed the complete explainability perspective and demonstrated that there are prerequisites to XAI success beyond new methods and techniques. I showed that explainability in expert contexts should be seen as a process that must be approached in stages. As a first step, I suggested procedures that should be followed to improve AI adoption and help users interact with new AI systems successfully. Then, I introduced design guidelines for expert-oriented explainability and showed how to design XAI interfaces to support expert decision-making. As a third step, I proposed design techniques to help ensure that experts engage with explainability, which would then have long-term value in increasing experts' AI literacy and domain knowledge. These contributions will help improve the usability of existing XAI methods and ensure that experts can fully benefit from explanations. This contribution is also generalisable to any domain beyond journalism and natural science. Instead of prescriptive guidelines, I provide examples and information that could help researchers and designers adapt and build their rules specific to the domain and context in which they work. More generally, this research will broaden researchers' and designers' perspectives and help them better understand domain experts' XAI needs.

7.1.1 Chapter 2 and 3 contributions

By conducting the XAI research literature review (Chapter 2), I provided a coherent overview of why explainability methods are often ineffective in expert contexts. This knowledge will allow XAI researchers to direct their attention to the gaps in the XAI research and will challenge their perspectives. By conducting the multidisciplinary review and extrapolating from HFE research literature (Chapter 3), I contributed to the HCI, AI, and XAI research knowledge. Based on the reviews, I provided a rich overview of expert-AI interaction challenges, including reasons for trust-related issues and experts' inability to stay in the decision-making loop. By extrapolating

from the HFE research, I expanded the knowledge base of human-AI interactions, filling some of the gaps in the HCI, AI, and XAI research and predicting future challenges and opportunities. By combining these extrapolations with the review of current XAI limitations, I also showed that XAI research needs to consider aspects such as cognitive workload, situation awareness, and workflow changes. The knowledge reported in this chapter will help shape expert-aware explainability and better tailor it to support a human-in-the-loop approach.

7.1.2 Chapter 4 contributions

This thesis also contributes to the scientific knowledge by demonstrating the practical challenges domain experts face when AI systems are introduced into their workflows and providing solutions for improved AI adoption (Chapter 4). I showed that AI, XAI, and HCI research communities often overlook aspects (e.g., early support, feedback-based training, and collaboration) that could determine whether experts will use AI tools and explanations. As a response, I proposed strategies to help experts effectively adopt and use AI systems and access explainability. These strategies could also improve work productivity and satisfaction and help developers and organisations use expert feedback to develop expert-aware technologies. More specifically, these findings will help the case study organisation and other organisations facing similar challenges to improve their AI implementation strategies and provide better support to domain experts. Findings reported in Chapter 4 also emphasised the importance of research approaches, such as participatory design, to improve AI and XAI contextual fit and usability.

7.1.3 Chapter 5 contributions

Based on the review of HFE research literature, I provided an informative overview demonstrating how expert decision-making strategies can be determined by the levels of users' expertise, contextual risk, and time-pressure (Chapter 5). I also outlined the critical decision-making strategies and cognitive biases linked to these aspects. Furthermore, the overview demonstrated why and how introducing intelligent systems can increase experts' cognitive workload and disrupt their decision-making. In Chapter 5, I also presented the ERT Framework, a domain-agnostic conceptual framework developed to inform explainability researchers and guide interface designers by providing information about critical decision-making

strategies to support any decision-making scenario. The framework provides a practical means of mapping significant contextual factors to appropriate explainability approaches and, in this way, introduces a pragmatic starting point for interface designers to rapidly incorporate explanations that are most likely to meet user needs and improve system understanding. Based on the workshops with UX designers and post-graduate design students, I also provided a worked example of AI in journalism and multiple design suggestions for various design goals in high-risk, high-expertise and high-time-pressure contexts. The ERT Framework will increase UX designers' awareness of the differences between experts' and novices' XAI needs. It will nudge them to think about the decision-making context differently, challenging them to expand their views. The framework will be helpful for UX designers and HCI, XAI, and AI researchers, as it provides a novel, user-centred approach to explainability in decision-making based on tailoring explanations to facilitate using experts' naturalistic decision-making and sensemaking strategies. Finally, the HFE literature review and introduction of the ERT Framework are the first steps towards understanding and acknowledging the changes AI introduces in expert decision-making. This work invites researchers and practitioners to recognise that these changes might disrupt decision-making and recognise how explainability could dampen or reverse the effect of this disruption.

7.1.4 Chapter 6 contributions

Lastly, in Chapter 6, I demonstrated the temporality issues of current explainability approaches. I outlined reasons why explanations must have a clear value to experts and should be engaging. I also showed how XAI could promote AI literacy and support expert skill development. Based on the Cognitive Psychology literature, I outlined design solutions that could inform researchers and practitioners on designing explainability that would be engaging and have long-lasting educational value. I suggested what design solutions could help to ensure that experts attended to explanations and proposed a framework that could inform when engaging explanations should be used, depending on the contextual factors impacting users' cognitive load and learning needs. This novel approach could help researchers recognise XAI's potential to enhance AI literacy and support skill development.

7.2 Future research

The next step will be to explore the applicability of the ERT Framework in practice. In future research, I will aim to provide design examples for all the dynamics, including low-risk and low-expertise contexts. This could be achieved by organising additional workshops with the UX designers working in various domains. In future research, I also aim to test the ERT Framework in practice. For example, by collaborating with UX designers and XAI developers, the ERT Framework could be used to develop an explainability interface, which would then be tested by interviewing and observing experts using it. It would be interesting to examine how UX designers and developers adopt the framework and modify the proposed design examples, what they take from it and how it affects their perspective on XAI for domain experts. Usability tests with experts using tailored XAI interfaces could help assess how this approach affects domain experts' ability to interact with AI and interpret its outputs.

As AI applications in journalism have expanded since the beginning of this project, it would be a valuable extension of this thesis to conduct ethnographically informed studies with media experts to understand how AI continue changing their decision-making in high-risk and high-time-pressure contexts. Ethnographic studies could also help to understand how media experts use explainability features and whether they do it effectively. It could also help to map their workflows more precisely, finding low-time-pressure moments when engaging explainability could be embedded and when experts could spend more time learning from them.

To further explore the research theme of cognitively engaging explanations, I aim to conduct studies examining the effects of proposed design strategies on domain knowledge and AI literacy. Future studies will also explore what other roles explainability could play in expert contexts. For example, they will test whether XAI could be designed to improve collaboration between experts and enable knowledge sharing. Lastly, I aim to conduct studies examining the perceived value of explanations and how this perception impacts experts' willingness to engage with them.

Appendices

Appendix A. The information sheet and consent form for the contextual inquiry and interview study with the software engineers, leadership members and experts at the life science institution.

THE UNIVERSITY of EDINBURGH
Edinburgh College of Art

BBC R&D

PETRAS

Information sheet for participants

Study title:	Explainability in AI-driven decision-making
Principal investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute

What is this document? This document explains the study that is being conducted, what your rights are, and what will be done with your data. You should keep this page for your records. After you read this, continue to the next page.

Background. AI-driven decision support systems are widely applied in various domains, such as healthcare, policing or natural sciences. These systems are fair and accountable. It is essential that humans can maintain meaningful agency and understand and oversee algorithmic processes. However, the introduction of these systems often means that experts are unable to use their expertise due to the disrupted flow of their decision-making. Often, experts do not adopt these systems, even when they could benefit from using them. This study explores challenges to effectively adopting AI tools into experts' workflow. Explainability is often seen as a promising mechanism for enabling human-in-the-loop, i.e., allowing a human expert to remain in the decision-making loop alongside AI tools. However, current approaches are ineffective and can lead to various biases. This study also explores how explainability could be tailored to support domain experts' naturalistic decision-making strategies, be more engaging, and promote learning.

Aims of the study. The study aims to explore i) the pre-introduction processes that create barriers to effective implementation of the software, ii) the key barriers preventing users from effectively adopting readily available AI-enabled software that they need for their projects, iii) how the views of experts and practitioners align when reflecting on the same processes, and how that relates to the low adoption of AI-

enabled software; and finally, iv) how teams collaborate and what role this collaboration play in improving effective adoption of AI systems.

Study agenda. Expert participants will be observed performing their usual workflow tasks and asked about their actions during the task (Contextual Inquiry Method). In cases where inquiring about participants' actions during the task is improbable, participants will be observed without interrupting experts (Direct Observation). Lastly, participants will be interviewed using a semi-structured interview method. Software engineers and members of the leadership will be only interviewed unless they see it useful to demonstrate aspects of the software.

The observations, demonstrations and interviews will be recorded, and notes will be taken. The information will only be kept until the data is analysed. Afterwards, the notes and recordings will be erased.

Risks and benefits. There are no known risks to participating in this study, and there are no tangible benefits to you. However, you will be contributing to our knowledge about the user-focused implementation of AI-driven systems.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with the Data Protection Law. To safeguard your privacy, we will never share personal information (like names or dates of birth) with anyone outside the research team; if you agree and want to be contacted for future studies, we will add your contact details to our secure participant database. Your data will be referred to by a unique participant number rather than by name. We will store any personal information using the University of Edinburgh's secure encrypted storage service or in a locked filing cabinet at the University of Edinburgh. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right to access can be exercised in accordance with the Data Protection Law. You also have other rights, including rights of correction, erasure and objection. For more details, including the right to complain with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during or after data gathering, we will delete your data, and there will be no penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you've just read, please feel free to ask or contact us later. You can contact the principal researcher by email at . This project has been approved by the ECA Ethics Committee.

Thank you for your help!

Participant consent and agreement to data usage

Study title:	Explainability in AI-driven decision-making
Principal investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute

PLEASE MARK EITHER 'YES' OR 'NO' FOR EVERY STATEMENT BELOW:

Consent for participation:	**Yes**	**No**
I consent to take part in the above study		

Agreement to identifiable data usage requests:	**Yes**	**No**
I agree that anonymised transcriptions of the contextual enquiries and interviews can be **shared with other researchers** and used for research purposes (e.g., presentations and publications).		

Participant name	Participant signature	Today's date

Unique participant code (researcher will complete)

Appendix B. The interview protocol for the interview study with the software engineers, leadership members and experts at the life science institution.

Principal Investigator: Auste Simkute

Interviewer: Auste Simkute

Introductions and hit record

I would like to speak to you about your experience with and usage of the in-house AI software* for data analysis. The goal is to understand how experts interact with new AI tools, how they form trust in these tools and how they benefit from their features, such as explainability. Also, how do their perceptions of these tools align or misalign with the perceptions of their developers? There are no right or wrong answers. This study does not evaluate you or your answers – we are interested in gathering a variety of people's experiences. Please stop me at any time and ask if anything is unclear.

*In this protocol, the in-house AI software is called the AI software or the in-house AI system. During the interview, it was called by its actual name, which was redacted for confidentiality purposes.

Interview questions for science experts

General questions and the use of the in-house AI software:

- Tell me more about your role and the key tasks you perform.
- Which, if any, of these tasks require the use of AI tools?
- How often do you use the in-house AI software?
- How would you describe how it works in your own words? (Is it AI or ML system, and why/how do you know? How do you understand AI/ML?)
- What are the other systems you use to support the same or similar tasks? What are the key differences between these systems?
- Could you tell me more about your preferences for one or the other system?
- How and when do you make a decision to trust the system? Its output?
- What aspects help you judge the trustworthiness of the system?
- Do you seek an explanation or any additional information after receiving an AI output? How do you do that?

- What do you expect from the in-house AI software and similar systems? What would be the ideal case scenario for it to be helpful in supporting your role?
- Can you tell me of a case/situation when the AI system was helpful?
- Can you tell me of a case/situation when the AI software frustrated you/didn't work as you expected? What did you do when that happened? How does the view or the way you use the system change after that?
- Do you plan to continue using AI software? How do you imagine its role in your work in the future?

Learning to use the software:

- Tell me more about the time this system was first introduced. The process of training, the first time you used it, and the way you progressed using the system.
- Have you noticed any changes in the way you use your expert knowledge/skills in this or similar tasks after using AI software?
- How has your understanding of the system changed over time? What really helped you to understand it better?
- What would you like to know about the system so that you can use it more efficiently/effectively? What would you change in the system in order to make it more interesting/enjoyable/easier to use?
- Tell me more about your preferred ways of learning about new technologies (including the in-house AI software) at your work.
- Where do you seek information/answers/clarifications when you run into an issue using the in-house AI software (or similar software)? What other sources would you like to have?

Between-team collaboration:

- How do you contact the AI team members? How often do you do that?
- Do you get approached by the AI team? What questions do they approach you with?
- Do the AI team members consult you about the software they develop?
- What is your experience working with software engineers and other AI team members?

- Could you share examples of times you collaborated with other teams when learning to use this or other software?

Interview questions for software developers

General questions and questions about developing the in-house AI software:

- Tell me more about your role and the key tasks you perform.
- What is your role regarding the in-house software?
- Tell me about the benefits of the in-house AI software for experts using it.
- What is the expert role in using in-house AI software? Why is it important?
- What would be an ideal expert-AI collaboration/working situation?
- What are the main challenges when it comes to the in-house AI software being implemented in an expert context?
- How have you been addressing these challenges/planning to address these challenges?
- What future challenges do you anticipate regarding the in-house AI software? How do you plan to address them?
- Did you consider factors such as level of expertise/time available/workflow when designing the system? If so, how?
- How do you support experts' ability to judge the trustworthiness of AI's outputs?
- Do you use explainability techniques to explain the in-house AI system and its outputs to experts? If so, please provide more information about your explainability strategy.

Training to use the software:

- What AI/ML background knowledge do you expect from experts using in-house AI software? What level of understanding about AI/ML/the specific software, etc., do they need to have to use it effectively? How do you measure or predict it?
- How do you introduce the new software to experts?
- What does the training process entail, and how is it developed?
- Do you contact experts after the training for feedback? What feedback do they give? Do they contact you with questions afterwards?

Between-team collaboration:

- What questions do you work with experts on? How do you contact them?
- Do you get approached by experts? What questions do they approach you with?
- Do you consult experts about the tools you develop? How often do you do that?
- How are experts involved in different stages of planning/development/implementation of the in-house AI software? What feedback do you get from them, and how do you respond? Do you inform them about the changes their feedback influences?
- What is your experience working with experts?
- Could you share examples of times you collaborated with other teams when learning to use this or other software?

Interview questions for the AI team lead and science director

General questions and questions about developing the in-house AI software:

- Tell me more about your role and the key tasks you perform.
- What are the general feelings within the expert teams about the in-house AI software?
- Have you noticed any patterns of experts accepting and rejecting it? What do you think about it?
- Are there any strategies you use to motivate experts to use/adopt this particular software and other in-house AI systems?
- How have expert teams' workflows changed since the in-house AI software was introduced?
- How have expert teams' dynamics changed since the in-house AI software was introduced?
- What would be an ideal way to use/implement these systems for an optimal/most effective and efficient outcome?
- What is the role of a human that needs to be augmented by the system in this decision-making/case?

- Have you noticed a change in how expert teams use their skills and make decisions after the introduction of in-house or other AI systems? What are your future predictions? How do you imagine decision-making in these roles will change in the future?
- What has the AI team struggled the most with/felt frustrated about after the introduction of the in-house AI tools?
- How could your team of engineers help experts to improve their decision-making effectiveness?

Training to use the software:

- How are the expert teams introduced to the in-house AI software?
- Are there any strategies you use to educate experts about these systems? Do you think the company's strategies are sufficient for the effective use of the in-house AI systems?
- What AI/ML background knowledge do you expect from experts using in-house AI software? What level of understanding about AI/ML/the specific software, etc., do they need to have to use it effectively? How do you measure or predict it?
- What does the training process entail, and how is it developed?
- Do you contact experts after the training for feedback? What feedback do they give? Do they contact you with questions afterwards?
- What are the future plans regarding the training?

Between-team collaboration:

- What questions do you work with expert teams on? How does your team contact experts?
- Do you or your team get approached by experts? What questions do they approach you with?
- Does your team consult experts about the tools you develop? How often do they do that?
- How are experts involved in different stages of planning/development/implementation of the in-house AI software? What

feedback do you get from them, and how do you respond? Do you inform them about the changes their feedback influences?

- Could you share examples of successful team collaboration/ projects where teams worked together?

Appendix C. Contextual inquiry and interview study details for ethics approval.

Research impact: The findings of the contextual inquiry study will help this and similar organisations to effectively introduce AI tools to their employees' workflows. It will help to understand the practical challenges that experts face when AI tools are introduced to support them and what support and training they need to be able to use them effectively. It will also help to understand the role explainability plays when experts have to use their expertise, as well as the outputs from the AI system. In a broader sense, the research outputs could inform future research AI adoption and invite the research community to look at explainability as part of the larger picture and approach it from a different perspective. Lastly, this research will provide motivation for conducting research in the *wild* or applying ethnographic methods when studying AI and its aspects (e.g., XAI) in expert contexts.

The research aims and questions

1. **Investigate the practical and contextual barriers that experts face when new AI systems are introduced into their workflows.**
    a. What practical difficulties do experts experience when new systems are introduced to support their work?
    b. How do experts respond to these practical difficulties? What strategies do they use to overcome them?
    c. What contextual factors influence experts' ability to use new AI tools?
    d. What strategies do experts use to respond to different contextual factors that interfere with or support their use of new AI tools?
2. **Identify the challenges to successful AI use and adoption occurring in different stages of AI system introduction, including before experts start using them.**
    a. How does the way the new AI system is introduced to experts affect their ability to adopt it and use it successfully?
    b. How does the training on using the new AI system experts receive affect their ability to adopt it and use it successfully?
    c. What factors interfere with experts' ability to use AI systems when they first try to use them independently?
    d. What factors interfere with experts' ability to use AI systems after they have already used them for a while?

3. **Outline the perceptions of the reasons for low in-house AI software adoption among experts, AI developers, and team leaders within the science organisation.**
    a. What do experts see as the key factors influencing their willingness to adopt new AI software?
    b. What do experts see as the key factors influencing their choice to reject new AI software or search for a replacement?
    c. What do AI developers see as the key factors influencing experts' willingness to adopt new AI software?
    d. What do experts see as the key factors influencing experts' choice to reject new AI software or search for a replacement?
    e. What do team leaders see as the key factors influencing experts' willingness to adopt new AI software?
    f. What do team leaders see as the key factors influencing experts' choice to reject new AI software or search for a replacement?
4. **Identify how different stakeholders' perceptions align or misalign and how that influences the adoption of an AI system.**
    a. How do these perceptions differ between the experts, team leaders and software developers?
    b. How do these differences affect the strategies team leaders use regarding AI adoption, including planning introduction, support and training?
    c. How do these differences affect the way software engineers interact with experts regarding the development of in-house AI systems?
    d. How do these differences affect the way experts interact with software engineers?
5. **Understand how AI developers, team leaders and experts approach the issue of low AI adoption.**
    a. What strategies do experts use to improve their ability to use in-house AI software more effectively?
    b. What strategies do team leaders use to increase experts' adoption of in-house software within and outside of their organisations?
    c. What strategies do software engineers use to increase in-house software adoption by experts within and outside of their organisations?

6. **Explore the role of collaboration between stakeholders for the AI adoption**
    a. How do experts interact with team leaders and software engineers?
    b. What effect do the presence and absence of these interactions have on experts' ability and willingness to adopt in-house AI tools?

Participants, consent, recruitment, compensation

Consent

Participants who agree to participate will be sent a digital copy of the consent form and information sheet to sign via email at least a day before their participation and will be asked to return it to the principal researcher.

Recruitment

The study will be conducted in a large natural science research centre whose developers develop in-house AI systems and which also houses the science experts who were recently introduced to these systems. However, the overall adoption of these systems among experts is low. The organisation was recognised as a potential case study during a public presentation of their work, in which they acknowledged the issues of AI adoption. After being approached by the principal researcher's supervisor, an informal meeting with the AI team leader was organised, and the invitation for the principal researcher to visit the organisation and conduct the study was made. The participants will be recruited via the AI team leader, who will inform the AI team members about the study and share the invitation to participate. The AI team leaders will also inform the other team leaders, who will share the study invitation with the experts in their teams.

Compensation

Participants will not be compensated for their participation due to the internal rules of the organisation in which the research was conducted.

Study user experience

Participants will be instructed to consent by signing a consent form that will be sent to them by email or given in person before their participation. After signing it, they will be asked to return it via email to the principal investigator. They will also be given a

physical copy of the consent form for their reference. They will receive the information sheet and the consent form at least one day before their participation. Interviews and observations will be conducted in person at the organisation during the scheduled times of the interviewees' working hours.
The interviews with team leaders will be conducted in their offices, the interviews with AI software engineers will be conducted in the booked conference room, and the interviews with experts will be conducted in the lab. They usually use the software or a booked conference room.
After consenting, participants will be briefed about the purpose of the study and their rights (5 min). Interviews will take about 60 min, and contextual inquiry tasks together with interviews will take about 90 min. Software developers and team leaders will be asked interview questions and allowed to demonstrate any software or information they deem important to support their answers. Experts will first be asked general questions and then asked to demonstrate the software they use and answer any questions that arise during the observation. Following the observation, the rest of the semi-structured interview will be conducted.

Data and measures

Participant's name. It will be collected after participants submit consent forms by emailing them or handing the physical copy to the principal investigator.
Data captured in notes. The notes will be taken throughout the observations in a physical notebook. These notes will include any information the researcher deems important to addressing the research aims and questions.
Audio recordings and transcripts. They will include the questions during the observations and semi-structured interviews.

Data handling, access, and storage

The identifiable contact data (name, email address) from the consent forms will be stored in a separate folder in a secure SharePoint site, accessible only to the project researchers. The consent forms will be retained for five years. Notes captured during the observations will be digitalised and stored in a separate, secure SharePoint location, restricted only to the project researchers. This data will be linked with an anonymous identifier unique to each participant. Any identifiable and confidential information will be edited to remove personally identifiable and confidential

information. After editing, the originals will be deleted, leaving only the edited version to be used as sources of narratives in publication. Some quotes from edited transcripts and questionnaire responses may be used verbatim and edited a second time to prevent identification. Anonymised audio recordings and transcripts will be stored in a separate, secure SharePoint location, restricted only to the project researchers. The principal investigator will control access to consent data, audio recordings, transcripts and notes. Anonymised study data will be retained for five years.

Voluntary participation and right to withdraw

Participants can withdraw from the study at any time and for any reason. If they choose to withdraw during or after data gathering, their data will be deleted, and there will be no penalty or loss of benefits to which they may be otherwise entitled.

Risk, mitigations, and benefits

Benefits. The potential benefit for the participants is that they can improve their practices. The team leader will be provided with a report on my observations that could improve the strategy of AI implementation and the training process of experts using AI systems. Experts could receive more support, information, and training that could improve their ability to use AI systems. In turn, this could benefit their data analysis process, making them more efficient and precise, which could lead to scientific discoveries in the area of natural science research. This study could also benefit other organisations that experience difficulties in effectively implementing AI software into their team workflows.

Risks. This study involves no risks. No evaluations will be made of individual participants' responses. Only aggregated and de-identified data will ever be reported. Participant responses in notes and recording transcripts will be de-identified to remove references to self, others, and any work content.

Use of study outputs

The outputs of the study will be used as a chapter of the Doctoral Thesis of Auste Simkute and as possible academic publication(s) for a journal such as the ACM Transaction on Computer-Human Interactions, or alternatively as conference proceedings for a conference such as ACM Conference on Fairness, Accountability, and Transparency.

Appendix D. The information sheet and consent form for the Pilot Design Workshop with postgraduate students.

THE UNIVERSITY of EDINBURGH
Edinburgh College of Art

BBC R&D

PETRAS

Information sheet for participants

Project title:	Explainability in AI and data-driven decision-making
Principal Investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute, Bronwyn Jones

What is this document? It explains the kind of study we're doing, your rights, and what will be done with your data. You should keep this page for your records. After you read this, continue to the next page.

Background. Algorithmic decision support systems are widely applied in domains ranging from healthcare to journalism. To ensure that these systems are fair and accountable, it is essential that humans can maintain meaningful agency and understand and oversee algorithmic processes. Explainability is often seen as a promising mechanism for enabling human-in-the-loop. However, current approaches are ineffective and can lead to various biases. This study argues that explainability should be tailored to support naturalistic decision-making and sensemaking strategies employed by domain experts and novices. Based on the Human Factors literature review, we recognised potential decision-making strategies dependent on expertise, risk and time dynamics. We propose the conceptual Expertise, Risk and Time Explainability framework (ERT), intended to be used as explainability design guidelines.

Aims of the workshop. The workshop aims to improve the ERT framework by recognising its aspects that might be challenging for designers, unfeasible, or require further development. It also aims to match proposed design goals with design approaches and examples. The workshop's outcomes will aid the understanding of how the proposed framework could be made more usable for design practitioners and decision-makers. Insights from this workshop will help populate the ERT

framework with examples of design strategies and will help shape informative design guidelines for explainability interface design.

Workshop agenda. The workshop will be held online on a ZOOM platform and last for 1,5 hours. During the workshop, participants will be introduced to the topic and will receive a briefing on the ERT design framework. Afterward participants will be split into two groups and asked to complete two ideation activities using a MIRO board. Finally, we will have a brief group discussion and you will have a chance to ask questions and give feedback.

Risks and benefits. There are no known risks to participating in this study, and there are no tangible benefits to you. However, you will be contributing to our knowledge about the ethics of data-driven project development.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with the Data Protection Law. To safeguard your privacy, we will never share personal information (like names) with anyone outside the research team. If you agree and want to be contacted for future studies, we will add your contact details to our secure participant database. Your data will be referred to by a unique participant number rather than by name. We will store any personal information using the University of Edinburgh's secure encrypted storage service or in a locked filing cabinet at the University of Edinburgh. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right to access can be exercised in accordance with the Data Protection Law. You also have other rights, including rights of correction, erasure and objection. For more details, including the right to complain with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during or after data gathering, we will delete your data, and there will be no penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you've just read, please feel free to ask or contact us later. The principal investigator can be contacted by email at . The ECA Ethics Committee has approved this project.

Thank you for your help!

Study title:	Explainability in AI and data-driven decision-making
Principal investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute, Bronwyn Jones

PLEASE MARK EITHER 'YES' OR 'NO' FOR EVERY STATEMENT BELOW:

Consent for participation:	**Yes**	**No**
I consent to take part in the above study		

Agreement to identifiable data usage requests:	**Yes**	**No**
I agree that anonymised transcriptions of the workshop can be **shared with other researchers** and used for research purposes (e.g., presentations and publications).		

Participant name	Participant signature	Today's date

Unique participant code (researcher will complete)

Appendix E. The information sheet and consent form for the Design Workshops with the BBC UX Designers.

Information sheet for participants

THE UNIVERSITY of EDINBURGH
Edinburgh College of Art

BBC R&D

PETRAS

Project title:	Explainability in AI and data-driven decision-making
Principal Investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute

What is this document? It explains the kind of study we're doing, your rights, and what will be done with your data. You should keep this page for your records. After you read this, continue to the next page.

Background. Algorithmic decision support systems are widely applied in domains ranging from healthcare to journalism. To ensure that these systems are fair and accountable, it is essential that humans can maintain meaningful agency and understand and oversee algorithmic processes. Explainability is often seen as a promising mechanism for enabling human-in-the-loop. However, current approaches are ineffective and can lead to various biases. This study argues that explainability should be tailored to support naturalistic decision-making and sensemaking strategies employed by domain experts and novices. Based on the Human Factors literature review, we recognised potential decision-making strategies dependent on expertise, risk and time dynamics. We propose the conceptual Expertise, Risk and Time Explainability framework (ERT), intended to be used as explainability design guidelines.

Aims of the workshop. The workshop aims to improve the ERT framework by recognising its aspects that might be challenging for designers, unfeasible, or require further development. It also aims to match proposed design goals with design approaches and examples. The workshop's outcomes will aid the understanding of how the proposed framework could be made more usable for design practitioners and decision-makers. Insights from this workshop will help populate the ERT framework with examples of design strategies and will help shape informative design guidelines for explainability interface design.

Workshop agenda. The workshop will be held online on a ZOOM platform. During the workshop, participants will be introduced to the topic, will receive a briefing on the ERT design framework and will be shown an example scenario of the framework being applied in practice. Afterwards, participants will be asked to complete an ideation activity using a MIRO board.

Risks and benefits. There are no known risks to participation in this study, likewise, there no tangible benefits to you, however you will be contributing to our knowledge about the ethics of data-driven project development.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with the Data Protection Law. To safeguard your privacy, we will never share personal information (like names) with anyone outside the research team. If you agree and want to be contacted for future studies, we will add your contact details to our secure participant database. Your data will be referred to by a unique participant number rather than by name. We will store any personal information using the University of Edinburgh's secure encrypted storage service or in a locked filing cabinet at the University of Edinburgh. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right to access can be exercised in accordance with the Data Protection Law. You also have other rights, including rights of correction, erasure and objection. For more details, including the right to complain with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during or after data gathering, we will delete your data, and there will be no penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you've just read, please feel free to ask or contact us later. The principal investigator can be contacted by email at . The ECA Ethics Committee has approved this project.

Thank you for your help!

Study title:	Explainability in AI and data-driven decision-making
Principal investigator:	Auste Simkute
Researchers collecting data:	Auste Simkute

PLEASE MARK EITHER 'YES' OR 'NO' FOR EVERY STATEMENT BELOW:

Consent for participation:	**Yes**	**No**
I consent to take part in the above study		

Agreement to identifiable data usage requests:	**Yes**	**No**
I agree that anonymised transcriptions of the workshop can be **shared with other researchers** and used for research purposes (e.g., presentations and publications).		

Participant name | Participant signature | Today's date

Unique participant code (researcher will complete)

Appendix F. Scenarios used for the Design Workshops.

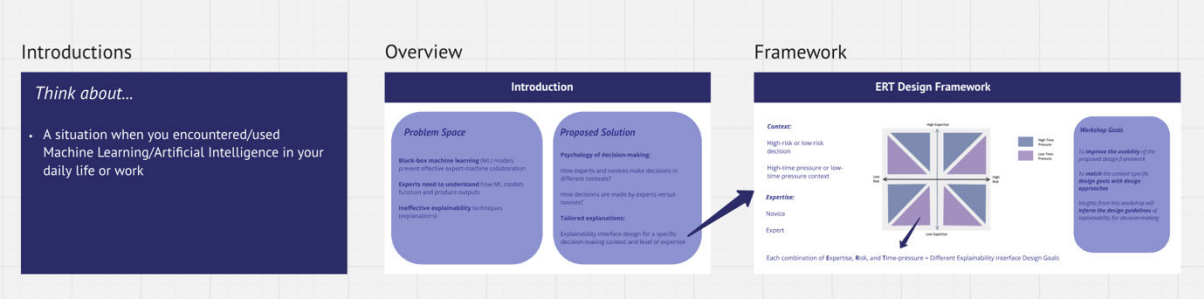
Social Work scenario (pilot study). The case worker, Brian, is a highly experienced social worker who has investigated multiple asylum requests and residence permit applications. Recently, he started using the ML system to support his decisions. The system just released an alert about one of the applications, prompting him to react to the case and investigate the application. However, Brian also sees that a similar application for asylum was fast-tracked for a rejection. The caseworker is confused about why this case needs to be investigated, but the other does not. The system does not explain why one case was prioritised and the other was rejected. Brian wishes an explanation was available, as he knows there is a high risk of an error that could disadvantage an applicant or even put their life in danger. Brian is not pressured by time to investigate the case he was alerted about. However, due to the high workload, he cannot re-investigate and question every decision/output of the ML system. Ben is part of a team designing an explainability interface for an AI system for processing asylum requests and residence permit applications. Social workers will use this system to see which requests should be prioritised and which need not be investigated further and can be rejected. They need to be able to justify their decision to their superiors and affected and otherwise involved stakeholders. They also have to take responsibility for any errors and trust that the system's suggestions are accurate and appropriate. Ben is trying to understand how to make this system and its outputs easier to understand for caseworkers, so he observes social worker Brian, who uses a system prototype which does not yet include an explanation component.

Journalism scenario. Ada is a highly experienced journalist. She is working on a breaking news story about the meeting of political leaders for a G7 summit. It is the day's top report that millions of people will read, so any errors could cause reputational damage to the news organisation and Ada herself (high-risk). Ada uses an AI-driven image recommendation tool, IRET, which extracts text from the article and uses entities such as names and locations to query the image library and pick pictures for journalists' stories. Ada types about a meeting between UK Prime Minister Boris Johnson and South African President Cyril Ramamphosa. IRET identifies keywords to suggest a 'top 5' selection of images based on the relevance and quality of the picture. Ada is unsure what Cyril Ramamphosa looks like and

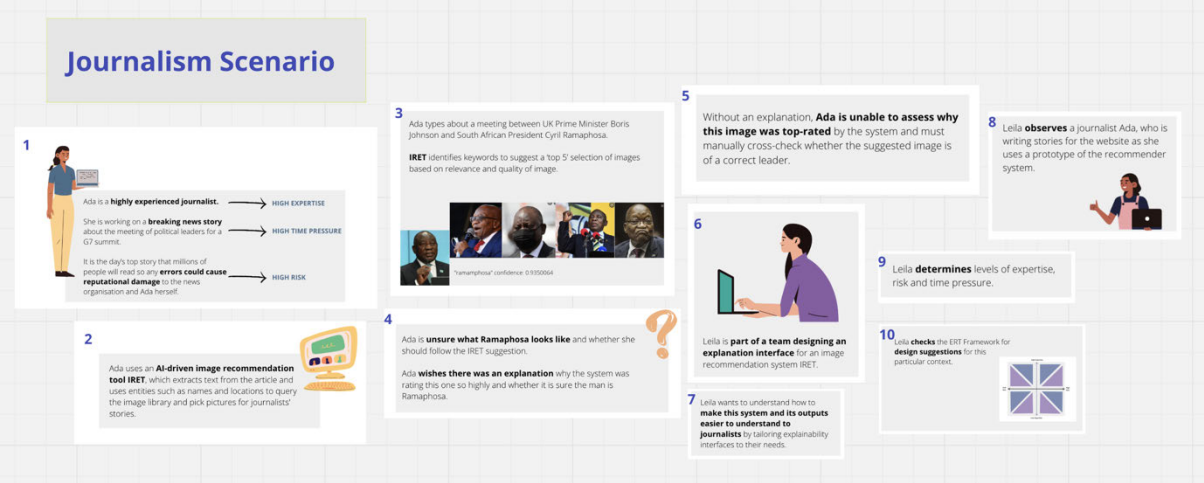
whether she should follow the IRET suggestion. Ada wishes there was an explanation as to why the system rated this image so highly and whether the man is actually Ramamphosa. Without an explanation, Ada cannot assess why this image was top-rated by the system and must manually cross-check whether the suggested image is of the correct leader. Leila is part of a team that is designing an explanation interface for an image recommendation system, IRET. Leila wants to understand how to make this system and its outputs easier to understand for journalists by tailoring explainability interfaces to their needs. Leila observes a journalist, Ada, who is writing stories for the website, and she uses a prototype of the recommender system. Leila determines levels of expertise, risk and time-pressure. Leila checks the ERT Framework for design suggestions for this particular context.

Appendix G. Design workshop task outline on Miro board.

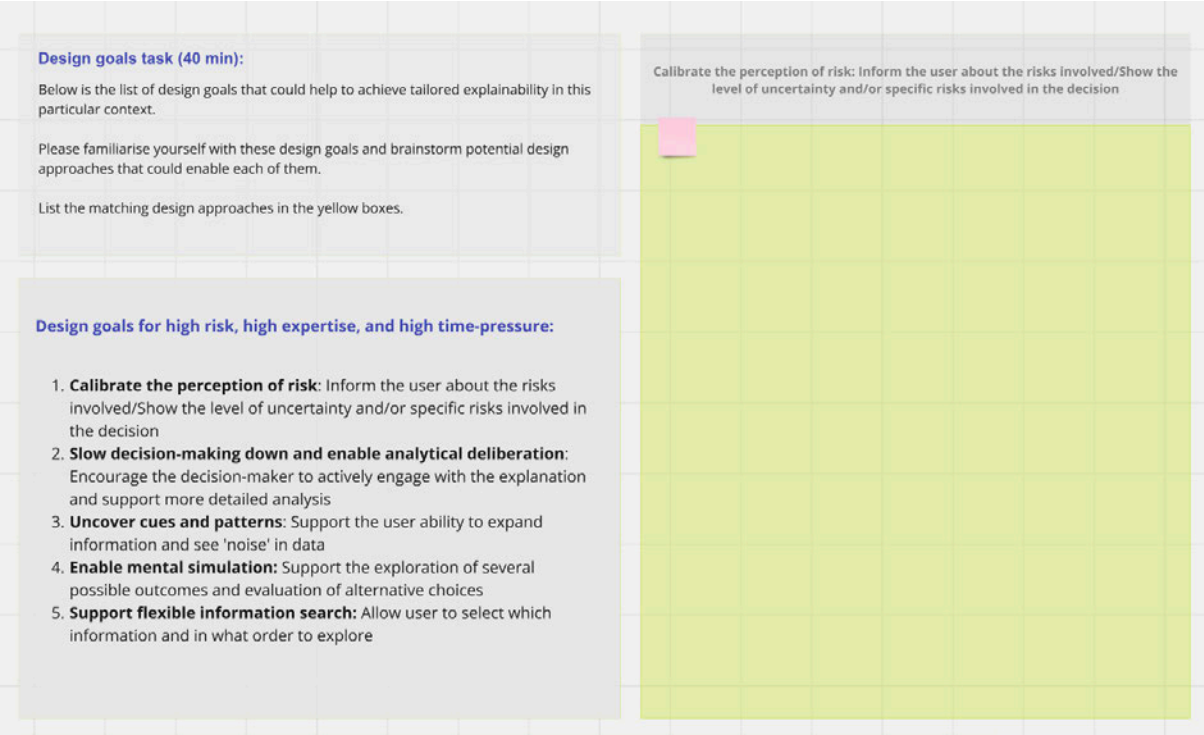
a) Workshop briefing and introduction of the ERT framework



b) The visual illustration of the Journalism scenario



c) The design goal task



Appendix H. Design workshop study details for ethics approval.

Research impact

The input from the design workshop will increase the usability and applicability of the ERT Framework and, in turn, make it a more effective informative tool for UX designers and AI and XAI researchers. The input from designers will help to populate the framework with exemplary design approaches that will allow anyone using the ERT Framework to shape their explainability interface design accordingly. The improved and design-feature-populated ERT Framework will be used to inform explainability interface design and help develop and design better explainability features for domain experts. This framework could help tailor explainability to more effectively support expert knowledge and AI output integration. The ERT guidelines are the first design guidelines for XAI in expert contexts. These guidelines are also the first to encourage designers to consider high-risk and time-pressure solutions and how to address high cognitive load in these contextual circumstances when designing explainability interfaces.

The research aims and questions

1. **Improve the ERT framework by recognising aspects that might be challenging for designers and aid the understanding of how the ERT framework could be made more usable for design practitioners and decision-makers**
    a) What aspects of the ERT framework are seen to be infeasible by designers?
    b) What aspects of the ERT framework are seen as the most useful by designers?
    c) What aspects of the ERT framework do designers struggle to envision applying in practice?
    d) Can designers recognise risk, time, and expertise dynamics? How could these dynamics be redefined to fit real-world scenarios from the designers' perspective?
    e) Are there any extra dynamics that should be considered besides the risk, time and expertise?

2. **Match proposed design goals with concrete design approaches, populate the ERT framework with examples of design strategies, and shape informative design guidelines for explainability interface design**
    a) How do designers envision the suggested design strategies from the UX perspective?
    b) Which design strategies do designers see as the most important?
    c) Which design strategies do designers see as the least important?
    d) What concrete examples of interface design features do designers match the suggested design strategies?

Participants, consent, recruitment, compensation

Consent

Participants who agree to participate will use an online consent form and study information sheet sent to them via email at least three days before the workshop.

Recruitment

For the pilot workshop, the principal researcher will recruit Postgraduate Design students from the University of Edinburgh. The principal researcher, Auste Simkute is a PhD student at the respective University and will undertake the recruitment using the Design department mailing list. BBC employees will be recruited using a snowballing effect by contacting leads and managers of various UX departments within the BBC and asking them to inform potential participants about this workshop.

Compensation

Pilot workshop participants will be compensated with a £10 bookshop gift voucher. The BBC employees will be compensated for their participation.

Study user experience

Participants will be instructed to consent via a consent form that will be sent to them by email at least three days before the workshop. After signing it, they will be asked to return it via email to the principal investigator. After consenting, they will be sent a Zoom meeting invitation for a set date and time. The workshop will take place via remote video call and will last approximately 90 minutes. Participants will be given Miro board access, where they will be introduced to the ERT framework and given an example with a visual illustration of how it could be applied in practice. Then,

participants will be asked to provide feedback on the applicability of the framework. Following these procedures, participants will be instructed to explore the five design goals and will be invited to ask the researcher clarifying questions. Afterwards, they will be given time to individually brainstorm design features to match the provided design guidelines from the ERT Framework using virtual sticky notes assigned to each goal. Lastly, participants will be invited to enable their microphones and contextualise, reason, explain and discuss design suggestions.

Data and measures

Participants' name and email address. Demographic information will be collected after participants submit consent forms by emailing them to the principal investigator. These details are needed to liaise with participants to schedule workshop times.

Data captured in digital notes. Notes will be taken throughout the workshops to capture participant explanations of their design suggestions, feedback about the framework, and other relevant responses to the study aims.

Data captured on the Miro board. The digital sticky notes where participants type their design suggestions will be saved as images.

Zoom recording of the workshops. Recordings will be used to clarify any responses that were captured in notes and on the Miro board.

Data handling, access, and storage

The identifiable contact data (name, email address) from the consent forms will be stored in a separate folder in a secure SharePoint site, accessible only to the project researchers. The consent forms will be retained for five years. The responses captured in notes and Miro boards will be stored in a separate, secure SharePoint location, restricted only to the project researchers. This data will be linked with an anonymous identifier unique to each participant. Any identifiable and confidential information will be edited to remove personally identifiable and confidential information. After editing, the originals will be deleted, leaving only the edited version to be used as sources of narratives in publication. Some quotes from edited transcripts and questionnaire responses may be used verbatim, edited a second time to prevent identification. Anonymised video and audio recordings will be stored in a separate, secure SharePoint location, restricted only to the project researchers.

The principal investigator will control access to consent data, recordings, screenshots and notes. Anonymised study data will be retained for 5 years.

Voluntary participation and right to withdraw

Participants will be able to withdraw from the study at any time and for any reason. If a participant chooses to withdraw from the study during or after data gathering, their data will be deleted, and there will be no penalty or loss of benefits to which they may be otherwise entitled.

Risk, mitigations, and benefits

Benefits. There are no direct benefits to participants. The principal researcher expects to learn about the UX designers' views of the ERT Framework and gather examples of design methods that could be used to populate the framework and contextualise the conceptual design guidelines. This study also has the potential to provide outputs that could support explainability researchers and UX designers at the BBC.

Risks. This study involves no risks. No evaluations will be made of individual participants' responses. Only aggregated and de-identified data will ever be reported. Participant responses in notes and screenshots of the Miro Board will be de-identified to remove references to self, others, and any work content.

Use of study outputs

The outputs of the study will be used as a chapter of the Doctoral Thesis of Auste Simkute and as possible academic publication(s) for a journal such as the Journal of Responsible Technology, or alternatively as conference proceedings for a conference such as ACM DIS: Designing Interactive Systems.

List of references

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3173574.3174156

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable Machine Learning in Healthcare. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 559–560. https://doi.org/10.1145/3233547.3233667

Ahmad, N. (2020). Refugees and Algorithmic Humanitarianism: Applying Artificial Intelligence to RSD Procedures and Immigration Decisions and Making Global Human Rights Obligations Relevant to AI Governance. *International Journal on Minority and Group Rights*, *1*(aop), 1–69. https://doi.org/10.1163/15718115-BJA10007

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. *Journal of Public Administration Research and Theory*, *33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

Alsheibani, S. A., Cheung, D. Y., & Messom, D. C. (n.d.). *Factors Inhibiting the Adoption of Artificial Intelligence at organizational-level: A Preliminary Investigation*.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*.

Altmann, E. M., Trafton, J. G., & Hambrick, D. Z. (2014). Momentary interruptions can derail the train of thought. *Journal of Experimental Psychology: General*, *143*(1), 215–226. https://doi.org/10.1037/a0030986

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *Bmc Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., Madai, V. I., & Initiative, on behalf of the Z.-I. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016. https://doi.org/10.1371/journal.pdig.0000016

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-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

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Anjomshoae, S., Främling, K., & Najjar, A. (2019). Explanations of Black-Box Model Predictions by Contextual Importance and Utility. In D. Calvaresi, A. Najjar, M.

Schumacher, & K. Främling (Eds.), *Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 95–109). Springer International Publishing. https://doi.org/10.1007/978-3-030-30391-4_6

Atoyan, H., Robert, J.-M., & Duquet, J.-R. (2008). Presentation of uncertain information in user interfaces to support decision making in complex military systems. *Proceedings of the 20th Conference on l'Interaction Homme-Machine*, 41–48. https://doi.org/10.1145/1512714.1512723

Ayres, L. B., Gomez, F. J. V., Linton, J. R., Silva, M. F., & Garcia, C. D. (2021). Taking the leap between analytical chemistry and artificial intelligence: A tutorial review. *Analytica Chimica Acta*, *1161*, 338403. https://doi.org/10.1016/j.aca.2021.338403

Badillo-Urquiola, K., Smriti, D., McNally, B., Golub, E., Bonsignore, E., & Wisniewski, P. J. (2019). Stranger Danger! Social Media App Features Co-designed with Children to Keep Them Safe Online. *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 394–406. https://doi.org/10.1145/3311927.3323133

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, *1*, 212–217 vol.1. https://doi.org/10.1109/ICSMC.2004.1398299

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Balzer, W. K., Doherty, M. E., & O'Connor Jr., R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*(3), 410–433. https://doi.org/10.1037/0033-2909.106.3.410

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). 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. https://doi.org/10.1145/3411764.3445717

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Barredo Arrieta, A., 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. https://doi.org/10.1016/j.inffus.2019.12.012

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *JOSA A*, *20*(7), 1391–1397. https://doi.org/10.1364/JOSAA.20.001391

Baudisch, P., Good, N., Bellotti, V., & Schraedley, P. (2002). 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. https://doi.org/10.1145/503376.503423

Baum, Z. J., Yu, X., Ayala, P. Y., Zhao, Y., Watkins, S. P., & Zhou, Q. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. *Journal of Chemical Information and Modeling*, *61*(7), 3197–3212. https://doi.org/10.1021/acs.jcim.1c00619

Baxter, G., Rooksby, J., Wang, Y., & Khajeh-Hosseini, A. (2012). The ironies of automation: Still going strong at 30? *Proceedings of the 30th European*

Conference on Cognitive Ergonomics, 65–71. https://doi.org/10.1145/2448136.2448149

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*. https://doi.org/10.1080/12460125.2021.1958505

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020). *Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach* (arXiv:2003.07703). arXiv. https://doi.org/10.48550/arXiv.2003.07703

Bekier, M., Molesworth, B. R. C., & Williamson, A. (2012). Tipping point: The narrow path between automation acceptance and rejection in air traffic management. *Safety Science*, *50*(2), 259–265. https://doi.org/10.1016/j.ssci.2011.08.059

Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, *4*. https://www.frontiersin.org/articles/10.3389/fdata.2021.688969

Benner, P., Tanner, C., & Chesla, C. (1992). From beginner to expert: Gaining a differentiated clinical world in critical care nursing. *Advances in Nursing Science*, *14*(3), 13.

Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society*, *28*(7), 806–822. https://doi.org/10.1080/10439463.2016.1253695

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). 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. https://doi.org/10.1145/3514094.3534164

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. https://doi.org/10.1145/3351095.3375624

Billings, C. E. (1991). Toward a Human-Centered Aircraft Automation Philosophy. *The International Journal of Aviation Psychology*, *1*(4), 261–270. https://doi.org/10.1207/s15327108ijap0104_1

Binmore, K. G., Kirman, A. P., & Tani, P. (1993). *Frontiers of Game Theory*. MIT Press.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159. https://proceedings.mlr.press/v81/binns18a.html

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics* (pp. 405–415). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_32

Birks, M., & Mills, J. (2022). *Grounded Theory: A Practical Guide*. SAGE.

Boch, S. J., Sezgin, E., & Lin, S. (2022). Ethical artificial intelligence in paediatrics. *The Lancet Child & Adolescent Health*, *6*. https://doi.org/10.1016/S2352-4642(22)00243-7

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-87480-9

Bohanec, M., Robnik-Sikonja, M., & Borstnar, M. K. (2017). Decision-making framework with double-loop learning through interpretable black-box machine

learning models. *Industrial Management & Data Systems*, *117*(7), 1389–1406. https://doi.org/10.1108/IMDS-09-2016-0409

Bolander, T. (2019). What do we loose when machines take the decisions? *Journal of Management & Governance*, *23*(4), 849–867. https://doi.org/10.1007/s10997-019-09493-x

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). 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. https://doi.org/10.1145/3490099.3511139

Brennen, A. (2020). 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. https://doi.org/10.1145/3334480.3383047

Breznitz, Z., Shaul, S., Horowitz-Kraus, T., Sela, I., Nevat, M., & Karni, A. (2013). Enhanced reading by training with imposed time constraint in typical and dyslexic adults. *Nature Communications*, *4*(1), Article 1. https://doi.org/10.1038/ncomms2488

Brooks, C., Gherhes, C., & Vorley, T. (2020). Artificial intelligence in the legal sector: Pressures and challenges of transformation. *Cambridge Journal of Regions, Economy and Society*, *13*(1), 135–152. https://doi.org/10.1093/cjres/rsz026

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

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

Bueff, A., Papantonis, I., Simkute, A., & Belle, V. (2022). *Explainability in Machine Learning: A Pedagogical Perspective* (arXiv:2202.10335). arXiv. https://doi.org/10.48550/arXiv.2202.10335

Bughin, J. R., Kretschmer, T., & Van Zeebroeck, N. (2019). Experimentation, Learning and Stress: The Role of Digital Technologies in Strategy Change. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3328421

Burgess, E. R., Jankovic, I., Austin, M., Cai, N., Kapuścińska, A., Currie, S., Overhage, J. M., Poole, E. S., & Kaye, J. (2023). Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3544548.3581251

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Burt, A., Leong, B., & Shirrell, S. (n.d.). *Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models*.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., & Terry, M. (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

Cao, H., Lee, C.-J., Iqbal, S., Czerwinski, M., Wong, P. N. Y., Rintel, S., Hecht, B., Teevan, J., & Yang, L. (2021). Large Scale Analysis of Multitasking Behavior During Remote Meetings. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445243

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., Via, A., & Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal*, *19*, 5762–5790. https://doi.org/10.1016/j.csbj.2021.10.009

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020). *The Emerging Landscape of Explainable AI Planning and Decision Making* (arXiv:2002.11697). arXiv. https://doi.org/10.48550/arXiv.2002.11697

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. https://doi.org/10.1002/stc.3019

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, *127*, 107018. https://doi.org/10.1016/j.chb.2021.107018

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 134–148. https://proceedings.mlr.press/v81/chouldechova18a.html

Christin, A. (2020). *Metrics at Work: Journalism and the Contested Meaning of Algorithms*. Princeton University Press.

Chun Tie, Y., Birks, M., & Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, *7*, 2050312118822927. https://doi.org/10.1177/2050312118822927

Clinciu, M. A., & Hastie, H. F. (2019). A Survey of Explainable AI Terminology. *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 8–13. https://doi.org/10.18653/v1/W19-8403

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education*, *37*(8), 695–703. https://doi.org/10.1046/j.1365-2923.2003.01577.x

Conejero, J. M., Preciado, J. C., Fernandez-Garcia, A. J., Prieto, A. E., & Rodriguez-Echeverria, 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. https://doi.org/10.1016/j.eswa.2020.114509

Cook, R., & Woods, D. (1997). Adapting to New Technology in the Operating Room. *Human Factors*, *38*, 593–613. https://doi.org/10.1518/001872096778827224

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Cork, R. D., Detmer, W. M., & Friedman, C. P. (1998). Development and Initial Validation of an Instrument to Measure Physicians' Use of, Knowledge about, and Attitudes Toward Computers. *Journal of the American Medical Informatics Association*, *5*(2), 164–176. https://doi.org/10.1136/jamia.1998.0050164

Cranefield, J., Winikoff, M., Chiu, Y.-T., Li, Y., Doyle, C., & Richter, A. (2023). Partnering with AI: The case of digital productivity assistants. *Journal of the Royal Society of New Zealand*, *53*(1), 95–118. https://doi.org/10.1080/03036758.2022.2114507

Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, *55*(1), 93–128.

Croskerry, P. (2003). The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them. *Academic Medicine*, *78*(8), 775.

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), Article 1. https://doi.org/10.1038/s41746-020-0254-2

Datta, A., Tschantz, M. C., & Datta, A. (2015). *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination* (arXiv:1408.6491). arXiv. https://doi.org/10.48550/arXiv.1408.6491

de Graaf, M., & Malle, B. (2017). *How people explain action (and AIS should too)*.

de Greeff, J., Jorritsma, W., & Neerincx, M. (n.d.). *The FATE System: FAir, Transparent and Explainable Decision Making*.

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

de Waard, D., van der Hulst, M., Hoedemaeker, M., & Brookhuis, K. A. (1999). Driver Behavior in an Emergency Situation in the Automated Highway System. *Transportation Human Factors*, *1*(1), 67–82. https://doi.org/10.1207/sthf0101_7

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). 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. https://doi.org/10.1145/3313831.3376638

Deng, L., Hu, Y., Cheung, J. P. Y., & Luk, K. D. K. (2017). A Data-Driven Decision Support System for Scoliosis Prognosis. *Ieee Access*, *5*, 7874–7884. https://doi.org/10.1109/ACCESS.2017.2696704

DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022a). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517441

DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022b). 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. https://doi.org/10.1145/3491102.3517441

Dey, S., Chakraborty, P., Kwon, B. C., Dhurandhar, A., Ghalwash, M., Suarez Saiz, F. J., Ng, K., Sow, D., Varshney, K. R., & Meyer, P. (2022). Human-centered explainability for life sciences, healthcare, and medical informatics. *Patterns*, *3*(5), 100493. https://doi.org/10.1016/j.patter.2022.100493

Dhurandhar, A., Iyengar, V., Luss, R., & Shanmugam, K. (2018). *TIP: Typifying the Interpretability of Procedures* (arXiv:1706.02952). arXiv. https://doi.org/10.48550/arXiv.1706.02952

Diab, D. L., Pui, S.-Y., Yankelevich, M., & Highhouse, S. (2011). Lay Perceptions of Selection Decision Aids in US and Non-US Samples. *International Journal of Selection and Assessment*, *19*(2), 209–216. https://doi.org/10.1111/j.1468-2389.2011.00548.x

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.

Diakopoulos, N. (2020). Computational News Discovery: Towards Design Considerations for Editorial Orientation Algorithms in Journalism. *Digital Journalism*, *8*(7), 945–967. https://doi.org/10.1080/21670811.2020.1736946

Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*, *5*(7), 809–828. https://doi.org/10.1080/21670811.2016.1208053

Dietvorst, B. J. (2016). *People Reject (Superior) Algorithms Because They Compare Them to Counter-Normative Reference Points* (SSRN Scholarly Paper 2881503). https://doi.org/10.2139/ssrn.2881503

Dietvorst, B. J., Simmons, J. P., & Massey, C. (n.d.). *Overcoming Algorithm Aversion:*

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), 114–126. https://doi.org/10.1037/xge0000033

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. https://doi.org/10.1016/j.ijhcs.2022.102792

Dobrow, M. J., Goel, V., Lemieux-Charles, L., & Black, N. A. (2006). The impact of context on evidence utilization: A framework for expert groups developing health policy recommendations. *Social Science & Medicine*, *63*(7), 1811–1824. https://doi.org/10.1016/j.socscimed.2006.04.020

Dodge, M. (2019). Women and White-Collar Crime. In *Oxford Research Encyclopedia of Criminology and Criminal Justice*. https://doi.org/10.1093/acrefore/9780190264079.013.493

Domeinski, J., Wagner, R., Schöbel, M., & Manzey, D. (2007). Human Redundancy in Automation Monitoring: Effects of Social Loafing and Social Compensation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(10), 587–591. https://doi.org/10.1177/154193120705101004

Dorton, S. L., Harper, S. B., & Neville, K. J. (2022). Adaptations to Trust Incidents with Artificial Intelligence. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 95–99. https://doi.org/10.1177/1071181322661146

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Dreyfus, H. L., & Dreyfus, S. E. (1986). From Socrates to Expert Systems: The Limits of Calculative Rationality. In C. Mitcham & A. Huning (Eds.), *Philosophy and Technology II: Information Technology and Computers in Theory and Practice* (pp. 111–130). Springer Netherlands. https://doi.org/10.1007/978-94-009-4512-8_9

Du, N., Huang, K. Y., & Yang, X. J. (2020). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, *62*(6), 987–1001. https://doi.org/10.1177/0018720819862916

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. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Dudai, Y., Karni, A., & Born, J. (2015). The Consolidation and Transformation of Memory. *Neuron*, *88*(1), 20–32. https://doi.org/10.1016/j.neuron.2015.09.004

Duhaime, I. M., & Schwenk, C. R. (1985). Conjectures on Cognitive Simplification in Acquisition and Divestment Decision Making. *Academy of Management Review*, *10*(2), 287–295. https://doi.org/10.5465/amr.1985.4278207

Edmondson, A. C., Bohmer, R. M., & Pisano, G. P. (2001). Disrupted Routines: Team Learning and New Technology Implementation in Hospitals. *Administrative Science Quarterly*, *46*(4), 685–716. https://doi.org/10.2307/3094828

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* (pp. 449–466). Springer International Publishing. https://doi.org/10.1007/978-3-030-60117-1_33

Ehsan, U., Saha, K., De Choudhury, M., & Riedl, M. O. (2023). Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–32. https://doi.org/10.1145/3579467

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing Transparency Design into Practice. *23rd International Conference on Intelligent User Interfaces*, 211–223. https://doi.org/10.1145/3172944.3172961

Eick, S. G., & Wills, G. J. (1995). High interaction graphics. *European Journal of Operational Research*, *81*(3), 445–459. https://doi.org/10.1016/0377-2217(94)00188-I

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral Decision Theory: Processes of Judgement and Choice. *Annual Review of Psychology*, *32*(1), 53–88. https://doi.org/10.1146/annurev.ps.32.020181.000413

Elwyn, G., Scholl, I., Tietbohl, C., Mann, M., Edwards, A. G., Clay, C., Légaré, F., Weijden, T. van der, 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), S14. https://doi.org/10.1186/1472-6947-13-S2-S14

Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (2023). Supporting Human-AI Teams:Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574. https://doi.org/10.1016/j.chb.2022.107574

European Parliament. Directorate General for Parliamentary Research Services. (2019). *Understanding algorithmic decision-making: Opportunities and challenges.* Publications Office. https://data.europa.eu/doi/10.2861/536131

Eva, K. W., & Cunnington, J. P. W. (2006). The difficulty with experience: Does practice increase susceptibility to premature closure? *Journal of Continuing Education in the Health Professions*, *26*(3), 192–198. https://doi.org/10.1002/chp.69

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, *19*(2), 133–144. https://doi.org/10.1177/15485129211028651

Flügge, A. A. (2021). 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. https://doi.org/10.1145/3462204.3481787

Frese, M., Brodbeck, F., Heinbokel, T., Mooser, C., Schleiffenbaum, E., & Thiemann, P. (1991). Errors in Training Computer Skills: On the Positive Function of Errors. *Human–Computer Interaction*, *6*(1), 77–93. https://doi.org/10.1207/s15327051hci0601_3

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), Article 1. https://doi.org/10.1038/s41746-021-00385-9

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadewar, S., Yang, Q., & Jahanshad, N. (2019). Towards human-guided machine learning. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 614–624. https://doi.org/10.1145/3301275.3302324

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

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. https://doi.org/10.1016/j.tips.2019.06.001

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

Gonzalez, C. (2005). Task Workload and Cognitive Abilities in Dynamic Decision Making. *Human Factors*, *47*(1), 92–101. https://doi.org/10.1518/0018720053653767

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, *38*(3), Article 3. https://doi.org/10.1609/aimag.v38i3.2741

Goodman, L. A. (1961). Snowball Sampling. *The Annals of Mathematical Statistics*, *32*(1), 148–170.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, *62*, 729–754. https://doi.org/10.1613/jair.1.11222

Green, B., & Chen, Y. (2019a). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Green, B., & Chen, Y. (2019b). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359152

Green, B., & Chen, Y. (2020). Algorithm-in-the-Loop Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), Article 09. https://doi.org/10.1609/aaai.v34i09.7115

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614. https://doi.org/10.1016/j.jsis.2020.101614

Grubb, P. L., Warm, J. S., Dember, W. N., & Berch, D. B. (1995). Effects of Multiple-Signal Discrimination on Vigilance Performance and Perceived Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *39*(21), 1360–1364. https://doi.org/10.1177/154193129503902101

Gu, D., Su, K., & Zhao, H. (2020). A case-based ensemble learning system for explainable breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, *107*, 101858. https://doi.org/10.1016/j.artmed.2020.101858

Gu, D., Zhao, W., Xie, Y., Wang, X., Su, K., & Zolotarev, O. V. (2021). A Personalized Medical Decision Support System Based on Explainable Machine Learning Algorithms and ECC Features: Data from the Real World. *Diagnostics*, *11*(9), Article 9. https://doi.org/10.3390/diagnostics11091677

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). *Local Rule-Based Explanations of Black Box Decision Systems* (arXiv:1805.10820). arXiv. https://doi.org/10.48550/arXiv.1805.10820

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), Article 2. https://doi.org/10.1609/aimag.v40i2.2850

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2014). Workload overload modeling: An experiment with MATB II to inform a computational model of task management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 849–853. https://doi.org/10.1177/1541931214581179

Hadash, S., Willemsen, M. C., Snijders, C., & IJsselsteijn, W. A. (2022). 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. https://doi.org/10.1145/3491102.3517650

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-

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. https://doi.org/10.1093/annonc/mdy166

Haldane, A. G., & May, R. M. (2011). Systemic risk in banking ecosystems. *Nature*, *469*(7330), Article 7330. https://doi.org/10.1038/nature09659

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, *69*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hansen, M., Roca-Sales, M., Keegan, J. M., & King, G. (2017). *Artificial Intelligence: Practice and Implications for Journalism*. https://doi.org/10.7916/D8X92PRD

Hanusch, F. (2019). Journalistic Roles and Everyday Life. *Journalism Studies*, *20*(2), 193–211. https://doi.org/10.1080/1461670X.2017.1370977

Hardavella, G., Aamli-Gaagnat, A., Saad, N., Rousalova, I., & Sreter, K. B. (2017). How to give and receive feedback effectively. *Breathe*, *13*(4), 327–333. https://doi.org/10.1183/20734735.009917

Hartikainen, M., Väänänen, K., Lehtiö, A., Ala-Luopa, S., & Olsson, T. (2022). Human-Centered AI Design in Reality: A Study of Developer Companies' Practices: A study of Developer Companies' Practices. *Nordic Human-Computer Interaction Conference*, 1–11. https://doi.org/10.1145/3546155.3546677

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456. https://doi.org/10.1016/j.clsr.2020.105456

Henry, K. E., Kornfield, R., Sridharan, A., Linton, R. C., Groh, C., Wang, T., Wu, A., Mutlu, B., & Saria, S. (2022). Human–machine teaming is key to AI adoption: Clinicians' experiences with a deployed machine learning system. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00597-7

Hilburn, B., Westin, C., & Borst, C. (2014). Will Controllers Accept a Machine That Thinks like They Think? The Role of Strategic Conformance in Decision Aiding Automation. *Air Traffic Control Quarterly*, *22*(2), 115–136. https://doi.org/10.2514/atcq.22.2.115

Hind, M., Wei, D., Campbell, M., Codella, N. C. F., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., & Varshney, K. R. (2019). TED: Teaching AI to Explain its Decisions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129. https://doi.org/10.1145/3306618.3314273

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., & Zuckerman, O. (2019). Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290605.3300645

Höferlin, B., Netzel, R., Höferlin, M., Weiskopf, D., & Heidemann, G. (2012). Inter-active learning of ad-hoc classifiers for video visual analytics. *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 23–32. https://doi.org/10.1109/VAST.2012.6400492

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. https://doi.org/10.1016/0030-5073(81)90040-4

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). Use of the Critical Decision Method to Elicit Expert Knowledge: A Case Study in the Methodology of Cognitive Task Analysis. *Human Factors*, *40*(2), 254–276. https://doi.org/10.1518/001872098779480442

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For "Explainable Ai". *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 197–201. https://doi.org/10.1177/1541931218621047

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. https://doi.org/10.1006/obhd.1995.1039

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, *106*, 106190. https://doi.org/10.1016/j.chb.2019.106190

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:1712.09923). arXiv. http://arxiv.org/abs/1712.09923

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. https://doi.org/10.1016/j.giq.2018.02.001

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://doi.org/10.1145/3479864

Hun Lee, M., Siewiorek, D. P., Smailagic, A., Bernardino, A., & Bermúdez 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*(2), 545–569. https://doi.org/10.1007/s11257-022-09348-5

Hung, T.-W., & Yen, C.-P. (2021). On the person-based predictive policing of AI. *Ethics and Information Technology*, *23*(3), 165–176. https://doi.org/10.1007/s10676-020-09539-x

Hutton, R. J. B., & Klein, G. (1999). Expert decision making. *Systems Engineering*, *2*(1), 32–45.

Inkpen, K., Chancellor, S., De Choudhury, M., Veale, M., & Baumer, E. P. S. (2019). Where is the Human?: Bridging the Gap Between AI and HCI. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290607.3299002

Jacobs, M., Pradier, M. F., McCoy, 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), Article 1. https://doi.org/10.1038/s41398-021-01224-x

Janssen, C. P., Brumby, D. P., Dowell, J., Chater, N., & Howes, A. (2011). Identifying Optimum Performance Trade-Offs Using a Cognitively Bounded Rational Analysis Model of Discretionary Task Interleaving. *Topics in Cognitive Science*, *3*(1), 123–139. https://doi.org/10.1111/j.1756-8765.2010.01125.x

Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019a). History and future of human-automation interaction. *International Journal of Human-Computer Studies*, *131*, 99–107. https://doi.org/10.1016/j.ijhcs.2019.05.006

Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019b). History and future of human-automation interaction. *International Journal of Human-Computer Studies*, *131*, 99–107. https://doi.org/10.1016/j.ijhcs.2019.05.006

Janssen, C. P., Gould, S. J. J., Li, S. Y. W., Brumby, D. P., & Cox, A. L. (2015). Integrating knowledge of multitasking and interruptions across different perspectives and research methods. *International Journal of Human-Computer Studies*, *79*, 1–5. https://doi.org/10.1016/j.ijhcs.2015.03.002

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. https://doi.org/10.1177/0894439320980118

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011

JMIR Medical Informatics—Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. (n.d.). Retrieved 4 September 2023, from https://medinform.jmir.org/2018/2/e24/

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250. https://doi.org/10.1073/pnas.1012933107

Johnson-Laird, P., & Shafir, E. (1993). The interaction between reasoning and decision making: An introduction. *Cognition*, *49*, 1–9. https://doi.org/10.1016/0010-0277(93)90033-R

Jones, A., & Petre, M. (1994). COMPUTER-BASED PRACTICAL WORK AT A DISTANCE: A CASE STUDY. In *Computer Assisted Learning: Selected*

Contributions from the CAL '93 Symposium (pp. 27–37). Elsevier. https://doi.org/10.1016/B978-0-08-041945-9.50012-3

Jones, B., & Luger, E. (2021). *AI and Journalism—Intelligible Cloud and Edge AI (ICE-AI)*. https://www.research.ed.ac.uk/en/publications/ai-and-journalism-intelligible-cloud-and-edge-ai-ice-ai

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine*, *67*(6), 507–512.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153. https://doi.org/10.1080/1463922021000054335

KAEMPF, G. L., KLEIN, G., THORDSEN, M. L., & WOLF, S. (1996). Decision Making in Complex Naval Command-and-Control Environments. *Human Factors*, *38*(2), 220–231. https://doi.org/10.1177/001872089606380204

Kahneman, D. (2006). A perspective on judgment and choice: Mapping bounded rationality. In *Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues.* Psychology Press.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kankanhalli, A., Charalabidis, Y., & Mellouli, S. (2019). IoT and AI for Smart Government: A Research Agenda. *Government Information Quarterly*, *36*(2), 304–309. https://doi.org/10.1016/j.giq.2019.02.003

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of

Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Keith, N., Richter, T., & Naumann, J. (2010). Active/Exploratory Training Promotes Transfer Even in Learners with Low Motivation and Cognitive Ability. *Applied Psychology*, *59*(1), 97–123. https://doi.org/10.1111/j.1464-0597.2009.00417.x

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review*, *103*(4), 687–719. https://doi.org/10.1037/0033-295X.103.4.687

Khairat, S., Marc, D., Crosby, W., & Al Sanousi, 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

Klayman, J. (1988). Chapter 4 On the How and Why (not) of Learning from Outcomes. In B. Brehmer & C. R. B. Joyce (Eds.), *Advances in Psychology* (Vol. 54, pp. 115–162). North-Holland. https://doi.org/10.1016/S0166-4115(08)62172-X

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), 97–122. https://doi.org/10.1016/0010-0277(93)90037-V

Klein, G. (2008). Naturalistic Decision Making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385

Klein, G. (2015a). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 164–168. https://doi.org/10.1016/j.jarmac.2015.07.001

Klein, G. (2015b). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 164–168. https://doi.org/10.1016/j.jarmac.2015.07.001

Klein, G. A. (n.d.). Intuition at work: Why developing your gut instincts will make you better at what you do. *(No Title)*. Retrieved 14 November 2023, from https://cir.nii.ac.jp/crid/1130282270485337856

Klein, G. A. (Ed.). (1993). *Decision making in action: Models and methods*. Ablex Pub.

Klein, G., Calderwood, R., & Clinton-Cirocco, A. (2010). Rapid Decision Making on the Fire Ground: The Original Study Plus a Postscript. *Journal of Cognitive Engineering and Decision Making*, *4*(3), 186–209. https://doi.org/10.1518/155534310X12844000801203

Klein, G., Moon, B., & Hoffman, R. (2006a). Making Sense of Sensemaking 2: A Macrocognitive Model. *Intelligent Systems, IEEE*, *21*, 88–92. https://doi.org/10.1109/MIS.2006.100

Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems*, *21*(4), 70–73. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2006.75

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. (2007). A data-frame theory of sensemaking. *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*, 113–155.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). 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. https://doi.org/10.1145/3290605.3300641

Koehler, D. J., & Harvey, N. (2008). *Blackwell Handbook of Judgment and Decision Making*. John Wiley & Sons.

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

Kulviwat, S., Bruner II, G. C., Kumar, A., Nasco, S. A., & Clark, T. (2007). Toward a unified theory of consumer acceptance technology. *Psychology & Marketing*, *24*(12), 1059–1084. https://doi.org/10.1002/mar.20196

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

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. https://doi.org/10.1016/j.telpol.2020.101976

Laganá, L., Oliver, T., Ainsworth, A., & Edwards, M. (2011). Enhancing computer self-efficacy and attitudes in multi-ethnic older adults: A randomised controlled study. *Ageing & Society*, *31*(6), 911–933. https://doi.org/10.1017/S0144686X10001340

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019). Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 59–67. https://doi.org/10.1609/hcomp.v7i1.5280

Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety*, *25*(10), 808–820. https://doi.org/10.1136/bmjqs-2015-004417

Langer, E. J. (n.d.). The psychology of control. *(No Title)*. Retrieved 30 November 2023, from https://cir.nii.ac.jp/crid/1130000796770039552

Langford, K., Kille, T., Lee, S.-Y., Zhang, Y., & Bates, P. R. (2022). "In automation we trust"—Australian air traffic controller perspectives of increasing automation in air traffic management. *Transport Policy*, *125*, 352–362. https://doi.org/10.1016/j.tranpol.2022.07.001

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, J. D., & Seppelt, B. D. (2009). Human Factors in Automation Design. In S. Y. Nof (Ed.), *Springer Handbook of Automation* (pp. 417–436). Springer. https://doi.org/10.1007/978-3-540-78831-7_25

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

Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (1993). Focussing in reasoning and decision making. *Cognition*, *49*(1), 37–66. https://doi.org/10.1016/0010-0277(93)90035-T

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*, 107539. https://doi.org/10.1016/j.chb.2022.107539

León, G. A., Chiou, E. K., & Wilkins, A. (2021). Accountability Increases Resource Sharing: Effects of Accountability on Human and AI System Performance. *International Journal of Human–Computer Interaction*, *37*(5), 434–444. https://doi.org/10.1080/10447318.2020.1824695

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376590

Liebhaber, M., Kobus, D., & Feher, B. (2002). *Studies of U.S. Navy Cues, Information Order, and Impact of Conflicting Data*. 76.

Lim, B. Y., Yang, Q., Abdul, A., & Wang, D. (2019). Why these Explanations? Selecting Intelligibility Types for Explanation Goals. *Los Angeles*.

Lindgren, I. (2023). Ironies of Public Service Automation – Bainbridge Revisited. *Proceedings of the 24th Annual International Conference on Digital Government Research*, 395–404. https://doi.org/10.1145/3598469.3598514

Lipshitz, R., Klein, G., Orasanu, J., & Salas, E. (2001). Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, *14*(5), 331–352. https://doi.org/10.1002/bdm.381

Lipshitz, R., & Strauss, O. (1997). Coping with Uncertainty: A Naturalistic Decision-Making Analysis. *Organizational Behavior and Human Decision Processes*, *69*(2), 149–163. https://doi.org/10.1006/obhd.1997.2679

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Lukyanenko, R., Castellanos, A., Samuel, B., Tremblay, M., & Maass, W. (2021). *Research Agenda for Basic Explainable AI*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), Article 1. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). *A Grounded Interaction Protocol for Explainable Artificial Intelligence* (arXiv:1903.02409). arXiv. http://arxiv.org/abs/1903.02409

Magerko, B., Lansing, E., Wray, B., Holt, L., Stensrud, B., & Arbor, A. (2005). *IMPROVING INTERACTIVE TRAINING THROUGH INDIVIDUALIZED CONTENT AND INCREASED ENGAGEMENT*.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Maiden, M. (2018). *Interactive Morphonology: Metaphony in Italy*. Routledge.

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), 229–236. https://doi.org/10.3163/1536-5050.99.3.010

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Marconi, F. (2020). 2. ENABLERS: THE AI TECHNOLOGIES DRIVING JOURNALISTIC CHANGE. In *2. ENABLERS: THE AI TECHNOLOGIES DRIVING JOURNALISTIC CHANGE* (pp. 55–128). Columbia University Press. https://doi.org/10.7312/marc19136-005

Mark, G., Gudith, D., & Klocke, U. (2008). The cost of interrupted work: More speed and stress. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 107–110. https://doi.org/10.1145/1357054.1357072

Mark, G., Voida, S., & Cardello, A. (2012). 'A pace not dictated by electrons': An empirical study of work without email. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 555–564. https://doi.org/10.1145/2207676.2207754

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* (SSRN Scholarly Paper 3290708). https://doi.org/10.2139/ssrn.3290708

Marx, V. (2013). The big challenges of big data. *Nature*, *498*(7453), Article 7453. https://doi.org/10.1038/498255a

Mattu, J. A., Jeff Larson,Lauren Kirchner,Surya. (n.d.). *Machine Bias*. ProPublica. Retrieved 14 November 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

McDonald, B., & Spaaij, R. (2021). Social Inclusion and Solidarity Building Through Sport for Recently Arrived Migrants and Refugees in Australia. In P. Liamputtong (Ed.), *Handbook of Social Inclusion* (pp. 1–15). Springer International Publishing. https://doi.org/10.1007/978-3-030-48277-0_102-1

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. https://doi.org/10.1080/10580530.2020.1849465

Metzger, U., & Parasuraman, R. (2001). The Role of the Air Traffic Controller in Future Air Traffic Management: An Empirical Study of Active Control versus Passive Monitoring. *Human Factors*, *43*(4), 519–528. https://doi.org/10.1518/001872001775870421

Metzger, U., & Parasuraman, R. (2005). Automation in Future Air Traffic Management: Effects of Decision Aid Reliability on Controller Performance and Mental Workload. *Human Factors*, *47*(1), 35–49. https://doi.org/10.1518/0018720053653802

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* (2021050005). Preprints. https://www.preprints.org/manuscript/202105.0005/v2

Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2019). To explain or not to explain: The effects of personal characteristics when explaining music recommendations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 397–407. https://doi.org/10.1145/3301275.3302313

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342. https://doi.org/10.1145/3593013.3594001

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Maestro, R. F. D. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE*, *15*(2), e0229596. https://doi.org/10.1371/journal.pone.0229596

Mishra, P., & Kereluik, K. (2011). *What 21st Century Learning? A review and a synthesis*. 3301–3312. https://www.learntechlib.org/primary/p/36828/

Misra, D., Avula, V., Wolk, D. M., Farag, H. A., Li, J., Mehta, Y. B., Sandhu, R., Karunakaran, B., Kethireddy, S., Zand, R., & Abedi, V. (2021). Early Detection of Septic Shock Onset Using Interpretable Machine Learners. *Journal of Clinical Medicine*, *10*(2), 301. https://doi.org/10.3390/jcm10020301

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299–313. https://doi.org/10.1037/a0014402

Montgomery, H., Lipshitz, R., & Brehmer, B. (2004). *How Professionals Make Decisions*. CRC Press.

Moray, N., Lootsteen, P., & Pajak, J. (1986). Acquisition of Process Control Skills. *IEEE Transactions on Systems, Man, and Cybernetics*, *16*(4), 497–504. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/TSMC.1986.289252

Mosier, K. L., Sethi, N., McCauley, S., Khoo, L., & Orasanu, J. M. (2007). What You Don’t Know Can Hurt You: Factors Impacting Diagnosis in the Automated

Cockpit. *Human Factors*, *49*(2), 300–310. https://doi.org/10.1518/001872007X312513

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. https://doi.org/10.1145/3351095.3372850

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N., & Ali, R. (2021). Nudging through Friction: An Approach for Calibrating Trust in Explainable AI. *2021 8th International Conference on Behavioral and Social Computing (BESC)*, 1–5. https://doi.org/10.1109/BESC53957.2021.9635271

Naiseh, M., Cemiloglu, D., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable Recommendations and Calibrated Trust: Two Systematic User Errors. *Computer*, *54*(10), 28–37. https://doi.org/10.1109/MC.2021.3076131

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

Norman, D. A. (1999). Affordance, conventions, and design. *Interactions*, *6*(3), 38–43. https://doi.org/10.1145/301153.301168

Nourani, M., King, J., & Ragan, E. (2020). 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. https://doi.org/10.1609/hcomp.v8i1.7469

Office, F. E. (2020). “Trustworthy and Explainable AI” Achieved Through Knowledge Graphs and Social Implementation. *FUJITSU Sci. Tech. J.*, *56*(1).

Oliva, R., & Sterman, J. D. (2001). Cutting Corners and Working Overtime: Quality Erosion in the Service Industry. *Management Science*, *47*(7), 894–914. https://doi.org/10.1287/mnsc.47.7.894.9807

Orasanu, J. M. (2010). Chapter 5—Flight Crew Decision-Making. In B. G. Kanki, R. L. Helmreich, & J. Anca (Eds.), *Crew Resource Management (Second Edition)* (pp. 147–179). Academic Press. https://doi.org/10.1016/B978-0-12-374946-8.10005-6

Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science*. https://doi.org/10.1287/orsc.11.4.404.14600

Pan, Y., Froese, F., Liu, N., Hu, Y., & Ye, M. (2022). The adoption of artificial intelligence in employee recruitment: The influence of contextual factors. *The International Journal of Human Resource Management*, *33*(6), 1125–1147. https://doi.org/10.1080/09585192.2021.1879206

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance Consequences of Automation-Induced ‘Complacency’. *The International Journal of Aviation Psychology*, *3*(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

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. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. https://doi.org/10.1109/3468.844354

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021). Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445304

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Payne, J. W., Bettman, J. R., Coupey, E., & Johnson, E. J. (1992). A constructive process view of decision making: Multiple strategies in judgment and choice. *Acta Psychologica*, *80*(1), 107–141. https://doi.org/10.1016/0001-6918(92)90043-D

Pérez, M., & Isaac, G. (2020). *Leveraging explainable machine learning to raise awareness among preadolescents about gender bias in supervised learning* [Master thesis, Universitat Politècnica de Catalunya]. https://upcommons.upc.edu/handle/2117/329448

Perlow, L. A., Okhuysen, G. A., & Repenning, N. P. (2002). The Speed Trap: Exploring the Relationship Between Decision Making and Temporal Context. *Academy of Management Journal*, *45*(5), 931–955. https://doi.org/10.5465/3069323

Pisano, G. P. (1996). Learning-before-doing in the development of new process technology. *Research Policy*, *25*(7), 1097–1119. https://doi.org/10.1016/S0048-7333(96)00896-7

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Povyakalo, A. A., Alberdi, E., Strigini, L., & Ayton, P. (2013). How to Discriminate between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography. *Medical Decision Making*, *33*(1), 98–107. https://doi.org/10.1177/0272989X12465490

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), 31:1-31:34. https://doi.org/10.1145/3577009

Raby, M., & Wickens, C. D. (1994). Strategic Workload Management and Decision Biases in Aviation. *The International Journal of Aviation Psychology*, *4*(3), 211–240. https://doi.org/10.1207/s15327108ijap0403_2

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). *The Algorithmic Automation Problem: Prediction, Triage, and Human Effort* (arXiv:1903.12220). arXiv. http://arxiv.org/abs/1903.12220

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping Business With Artificial Intelligence: Closing the Gap Between Ambition and Action. *MIT Sloan Management Review*, *59*(1), n/a-0.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 19–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-98131-4_2

Rasmussen, J., & Vicente, K. J. (1989). Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies*, *31*(5), 517–534. https://doi.org/10.1016/0020-7373(89)90014-X

Raven, M. E., & Flanders, A. (1996). Using contextual inquiry to learn about your audiences. *ACM SIGDOC Asterisk Journal of Computer Documentation*, *20*(1), 1–13. https://doi.org/10.1145/227614.227615

Rayo, M. F., & Moffatt-Bruce, S. D. (2015). Alarm system management: Evidence-based guidance encouraging direct measurement of informativeness to improve alarm response. *BMJ Quality & Safety*, *24*(4), 282–286. https://doi.org/10.1136/bmjqs-2014-003373

Reason, J., Broadbent, D. E., Baddeley, A. D., & Reason, J. (1997). The contribution of latent human failures to the breakdown of complex systems. *Philosophical*

Transactions of the Royal Society of London. B, Biological Sciences, *327*(1241), 475–484. https://doi.org/10.1098/rstb.1990.0090

Reis, J., Santo, P. E., & Melao, N. (2019). Impacts of Artificial Intelligence on Public Administration: A Systematic Literature Review. *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–7. https://doi.org/10.23919/CISTI.2019.8760893

Renz, A., & Hilbig, R. (2020). Prerequisites for artificial intelligence in further education: Identification of drivers, barriers, and business models of educational technology companies. *International Journal of Educational Technology in Higher Education*, *17*(1), 14. https://doi.org/10.1186/s41239-020-00193-3

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.11491

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). *Beyond Accuracy: Behavioral Testing of NLP models with CheckList* (arXiv:2005.04118). arXiv. https://doi.org/10.48550/arXiv.2005.04118

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*(8), 1143–1146. https://doi.org/10.1007/s00787-021-01836-0

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Ross, K. G., Shafer, J. L., & Klein, G. (2006). Professional Judgments and "Naturalistic Decision Making". In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 403–420). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.023

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Rudisill, D. M. (1995). Line Pilots' Attitudes About And Experience With Flight Deck Automation: Results Of An International Survey And Proposed Guidelines. *Proceedings of the Eighth International Symposium on Aviation Psychology*.

Rundmo, T. (2001). Employee images of risk. *Journal of Risk Research*, *4*(4), 393–404. https://doi.org/10.1080/136698701100653259

Rundo, L., Militello, C., Vitabile, S., Russo, G., Sala, E., & Gilardi, M. C. (2020). A Survey on Nature-Inspired Medical Image Analysis: A Step Further in Biomedical Data Integration. *Fundamenta Informaticae*, *171*(1–4), 345–365. https://doi.org/10.3233/FI-2020-1887

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. https://doi.org/10.1016/j.neucom.2017.01.105

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. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/TSMC.1981.4308761

Salas, E., Wilson, K. A., Burke, C. S., & Bowers, C. A. (2002). Myths About Crew Resource Management Training. *Ergonomics in Design*, *10*(4), 20–24. https://doi.org/10.1177/106480460201000406

Samek, W., & Müller, K.-R. (2019). *Towards Explainable Artificial Intelligence* (Vol. 11700, pp. 5–22). https://doi.org/10.1007/978-3-030-28954-6_1

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: Effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. https://doi.org/10.1080/00140139.2015.1094577

Saxena, D., Badillo-Urquiola, K., Wisniewski, P. J., & Guha, S. (2020). A Human-Centered Review of Algorithms used within the U.S. Child Welfare System. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376229

Schaekermann, M., Cai, C. J., Huang, A. E., & Sayres, R. (2020). 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. https://doi.org/10.1145/3313831.3376290

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 240–251. https://doi.org/10.1145/3301275.3302308

Schemmer, M., Kühl, N., & Satzger, G. (2021). *Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Work Through Explainable Artificial Intelligence* (arXiv:2109.13827). arXiv. https://doi.org/10.48550/arXiv.2109.13827

Schmitt, J., & Klein, G. (1999). *A Recognitional Planning Model:* Defense Technical Information Center. https://doi.org/10.21236/ADA461179

Schnackenberg, A. K., & Tomlinson, E. C. (2016). Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships. *Journal of Management*, *42*(7), 1784–1810. https://doi.org/10.1177/0149206314525202

Schneeberger, T., Gebhard, P., Baur, T., & André, E. (2019). PARLEY: A transparent virtual social agent training interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, 35–36. https://doi.org/10.1145/3308557.3308674

Schneider, W. (1985). Training High-Performance Skills: Fallacies and Guidelines. *Human Factors*, *27*(3), 285–300. https://doi.org/10.1177/001872088502700305

Schoonderwoerd, T. A. J., 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. https://doi.org/10.1016/j.ijhcs.2021.102684

Schunk, D. H. (2012). Social cognitive theory. In *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues* (pp. 101–123). American Psychological Association. https://doi.org/10.1037/13273-005

Seagull, F. J., Wickens, C. D., & Loeb, R. G. (2001). When is Less More? Attention and Workload in Auditory, Visual, and Redundant Patient-Monitoring Conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(18), 1395–1399. https://doi.org/10.1177/154193120104501817

Sears, A., & Jacko, J. A. (2009). *Human-Computer Interaction: Development Process*. CRC Press.

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, T., De Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Selbst, A. D., & Barocas, S. (2018). THE INTUITIVE APPEAL OF EXPLAINABLE MACHINES. *Fordham Law Review*, *87*(3), 1085–1139.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Sellwood, M. A., Ahmed, M., Segler, M. H., & Brown, N. (2018). Artificial intelligence in drug discovery. *Future Medicinal Chemistry*, *10*(17), 2025–2028. https://doi.org/10.4155/fmc-2018-0212

Shani, C., Zarecki, J., & Shahaf, D. (2023). The Lean Data Scientist: Recent Advances Toward Overcoming the Data Bottleneck. *Communications of the ACM*, *66*(2), 92–102. https://doi.org/10.1145/3551635

Sheh, R. K.-M. (2017). ‘ Why Did You Do That?’ Explainable Intelligent Robots. *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. https://cdn.aaai.org/ocs/ws/ws0359/15162-68408-1-PB.pdf

Sheridan, T. B. (2012). Human Supervisory Control. In *Handbook of Human Factors and Ergonomics* (pp. 990–1015). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch34

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Sibbald, M., de Bruin, A. B. H., & van Merrienboer, J. J. G. (2013). Checklists improve experts’ diagnostic decisions. *Medical Education*, *47*(3), 301–308. https://doi.org/10.1111/medu.12080

Simkute, A., Luger, E., Evans, M., & Jones, R. (2020). Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context. *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 263–268. https://doi.org/10.1145/3393914.3395862

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*, *7–8*, 100017. https://doi.org/10.1016/j.jrt.2021.100017

Simkute, A., Surana, A., Luger, E., Evans, M., & Jones, R. (2022). 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. https://doi.org/10.1145/3547522.3547678

Simonson, I., Carmon, Z., & O’Curry, S. (1994). Experimental Evidence on the Negative Effect of Product Features and Sales Promotions on Brand Choice. *Marketing Science*, *13*(1), 23–40. https://doi.org/10.1287/mksc.13.1.23

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3544548.3581075

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. https://doi.org/10.1145/3375627.3375830

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health Psychology*, *24*(4, Suppl), S35–S40. https://doi.org/10.1037/0278-6133.24.4.S35

Smith, H. P. R. (1979). *A simulator study of the interaction of pilot workload with errors, vigilance, and decisions* (NASA-TM-78482). NASA. https://ntrs.nasa.gov/citations/19790006598

Smith, K., Shanteau, J., & Johnson, P. (2004). *Psychological Investigations of Competence in Decision Making*. Cambridge University Press.

Smith-Renner, A., Kleanthous, S., Lim, B., Kuflik, T., Stumpf, S., Otterbacher, J., Sarkar, A., Dugan, C., & Shulner, A. (2020). ExSS-ATEC: Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies 2020. *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (Iui'20)*, 7–8. https://doi.org/10.1145/3379336.3379361

Sousa, W. G. de, Melo, E. R. P. de, 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. https://doi.org/10.1016/j.giq.2019.07.004

Sterman, J. D. (n.d.). *Learning In and About Complex Systems*.

Stray, J. (2021). Making Artificial Intelligence Work for Investigative Journalism. In *Algorithms, Automation, and News*. Routledge.

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. https://doi.org/10.1175/BAMS-88-12-1893

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance*, *23*(1), 86–112. https://doi.org/10.1016/0030-5073(79)90048-5

Syed, R., Suriadi, S., Adams, M., Bandara, W., Leemans, S. J. J., Ouyang, C., Ter Hofstede, A. H. M., Van De Weerd, I., Wynn, M. T., & Reijers, H. A. (2020). Robotic Process Automation: Contemporary themes and challenges.

Computers in Industry, *115*, 103162. https://doi.org/10.1016/j.compind.2019.103162

Taekman, J. M., & Shelley, K. (2010). Virtual Environments in Healthcare: Immersion, Disruption, and Flow. *International Anesthesiology Clinics*, *48*(3), 101. https://doi.org/10.1097/AIA.0b013e3181eace73

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017). Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 1–6. https://doi.org/10.1145/3077257.3077260

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). *Investigating Human + Machine Complementarity for Recidivism Predictions* (arXiv:1808.09123). arXiv. https://doi.org/10.48550/arXiv.1808.09123

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems* (arXiv:1806.07552). arXiv. https://doi.org/10.48550/arXiv.1806.07552

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Urueña, S. (2019). Understanding “plausibility”: A relational approach to the anticipatory heuristics of future scenarios. *Futures*, *111*, 15–25. https://doi.org/10.1016/j.futures.2019.05.002

Vaccaro, K., Huang, D., Eslami, M., Sandvig, C., Hamilton, K., & Karahalios, K. (2018). The Illusion of Control: Placebo Effects of Control Settings. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173590

van Baalen, S., Boon, M., & Verhoef, P. (2021). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice*, *27*(3), 520–528. https://doi.org/10.1111/jep.13541

van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 28:1-28:21. https://doi.org/10.1145/3359130

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. https://doi.org/10.1016/j.artint.2020.103404

VanBerlo, B. (2021, June 2). *An open-source interpretable machine learning approach to prediction of chronic homelessness*. Medium. https://towardsdatascience.com/an-open-source-interpretable-machine-learning-approach-to-prediction-of-chronic-homelessness-8215707aa572

Veale, M., Van Kleek, M., & Binns, R. (2018). 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. https://doi.org/10.1145/3173574.3174014

Vedapradha, R., Hariharan, R., & Shivakami, R. (2019). Artificial Intelligence: A Technological Prototype in Recruitment. *Journal of Service Science and Management*, *12*(3), Article 3. https://doi.org/10.4236/jssm.2019.123026

Veitch, E., & Andreas Alsos, O. (2022). A systematic review of human-AI interaction in autonomous ship systems. *Safety Science*, *152*, 105778. https://doi.org/10.1016/j.ssci.2022.105778

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 186–204.

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

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

Wang, D., Weisz, J. D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., & Gray, A. (2019). Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the*

ACM on Human-Computer Interaction, *3*(CSCW), 1–24. https://doi.org/10.1145/3359313

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300831

Wang, X., & Yin, M. (2021). 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. https://doi.org/10.1145/3397481.3450650

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Wearing, M. M. O., Jim McLennan, Alexander J. (2004). How Expertise Is Applied in Real-World Dynamic Environments: Head-Mounted Video and Cued Recall as a Methodology for Studying Routines of Decision Making. In *The Routines of Decision Making*. Psychology Press.

Weber, R. O., Johs, A. J., Li, J., & Huang, K. (2018). Investigating Textual Case-Based XAI. In M. T. Cox, P. Funk, & S. Begum (Eds.), *Case-Based Reasoning Research and Development* (pp. 431–447). Springer International Publishing. https://doi.org/10.1007/978-3-030-01081-2_29

Weld, D. S., & Bansal, G. (2018). *The Challenge of Crafting Intelligible Intelligence* (arXiv:1803.04263). arXiv. http://arxiv.org/abs/1803.04263

Westin, C., Borst, C., & Hilburn, B. (2016). Strategic Conformance: Overcoming Acceptance Issues of Decision Aiding Automation? *IEEE Transactions on*

Human-Machine Systems, *46*(1), 41–52. IEEE Transactions on Human-Machine Systems. https://doi.org/10.1109/THMS.2015.2482480

Whalen, J. (1995). *Expert Systems versus Systems for Experts*.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Wickens, C. D., Santamaria, A., & Sebok, A. (2013). A Computational Model of Task Overload Management and Task Switching. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 763–767. https://doi.org/10.1177/1541931213571167

WIENER, E. L., & CURRY, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics*, *23*(10), 995–1011. https://doi.org/10.1080/00140138008924809

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. https://doi.org/10.1145/3351095.3372833

Williams, D. J., & Noyes, J. M. (2007). How does our perception of risk influence decision-making? Implications for the design of risk information. *Theoretical Issues in Ergonomics Science*, *8*(1), 1–35. https://doi.org/10.1080/14639220500484419

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. https://doi.org/10.1080/01900692.2018.1498103

Wolf, C. T. (2019). Explainability scenarios: Towards scenario-based XAI design. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 252–257. https://doi.org/10.1145/3301275.3302317

Woodruff, M. C., Ramonell, R. P., Nguyen, D. C., Cashman, K. S., Saini, A. S., Haddad, N. S., Ley, A. M., Kyu, S., Howell, J. C., Ozturk, T., Lee, S., Suryadevara, N., Case, J. B., Bugrovsky, R., Chen, W., Estrada, J., Morrison-Porter, A., Derrico, A., Anam, F. A., … Sanz, I. (2020). Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nature Immunology*, *21*(12), Article 12. https://doi.org/10.1038/s41590-020-00814-z

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. https://doi.org/10.1016/j.artint.2022.103839

Yang, C.-L., Yuan, C. W. (Tina), & Wang, H.-C. (2019). When Knowledge Network is Social Network: Understanding Collaborative Knowledge Transfer in Workplace. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359266

Yang, K., & Stoyanovich, J. (2017). Measuring Fairness in Ranked Outputs. *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 1–6. https://doi.org/10.1145/3085504.3085526

Yang, L., Wang, H., & Deleris, L. A. (2021). What Does It Mean to Explain? A User-Centered Study on AI Explainability. In H. Degen & S. Ntoa (Eds.), *Artificial*

Intelligence in HCI (pp. 107–121). Springer International Publishing. https://doi.org/10.1007/978-3-030-77772-2_8

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

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). 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. https://doi.org/10.1145/3290605.3300509

Young, M., Rodriguez, L., Keller, E., Sun, F., Sa, B., Whittington, J., & Howe, B. (2019). Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 191–200. https://doi.org/10.1145/3287560.3287577

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User Trust Dynamics: An Investigation Driven by Differences in System Performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3025171.3025219

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

Zhang, A. Y., Lam, S. S. W., Liu, N., Pang, Y., Chan, L. L., & Tang, P. H. (2018). Development of a Radiology Decision Support System for the Classification of MRI Brain Scans. *2018 Ieee/Acm 5th International Conference on Big Data Computing Applications and Technologies (Bdcat)*, 107–115. https://doi.org/10.1109/BDCAT.2018.00021

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, *2*, 100025. https://doi.org/10.1016/j.caeai.2021.100025

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). 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. https://doi.org/10.1145/3351095.3372852

Zhao, Q., Xu, H., Li, J., Rajput, F. A., & Qiao, L. (2024). The Application of Artificial Intelligence in Alzheimer's Research. *Tsinghua Science and Technology*, *29*(1), 13–33. Tsinghua Science and Technology. https://doi.org/10.26599/TST.2023.9010037

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & 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

Zhou, Y., Li, Z., & Li, Y. (2021). Interdisciplinary collaboration between nursing and engineering in health care: A scoping review. *International Journal of Nursing Studies*, *117*, 103900. https://doi.org/10.1016/j.ijnurstu.2021.103900

Zytko, D., J. Wisniewski, P., Guha, S., P. S. Baumer, E., & Lee, M. K. (2022). Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. *CHI Conference on*

Human Factors in Computing Systems Extended Abstracts, 1–4.

https://doi.org/10.1145/3491101.3516506