Explainability for experts

Journal of Responsible Technology 7-8 (2021) 100017

Contents lists available at ScienceDirect

Journal of Responsible Technology

journal homepage: www.sciencedirect.com/journal/journal-of-responsible-technology

Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable

Auste Simkute [a,*], Ewa Luger [a], Bronwyn Jones [a], Michael Evans [b], Rhianne Jones [b]

[a] University of Edinburgh, Edinburgh, United Kingdom
[b] BBC Research and Development, Salford, Greater Manchester, United Kingdom

ARTICLE INFO

Keywords:
Explainability
Decision support systems
Journalism
Human-in-the-loop
Expertise

ABSTRACT

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, 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. We argue that explainability should be tailored to support naturalistic decision-making and sensemaking strategies employed by domain experts and novices. Based on cognitive psychology and human factors literature review we map potential decision-making strategies dependant on expertise, risk and time dynamics and propose the conceptual Expertise, Risk and Time Explainability framework, intended to be used as explainability design guidelines. Finally, we present a worked example in journalism to illustrate the applicability of our framework in practice.

1. Introduction

A growing number of domain experts find themselves having to rely on Artificial Intelligence (AI) or Machine Learning (ML) systems' generated risk assessment scores, predictions, or other types of algorithmic outputs when making decisions. Ensuring that expert users can understand, oversee, supervise, and control the process of algorithmic decision-making is essential, as Decision Support Systems (DSS) are being increasingly deployed to support decision-making in domains that are socio-technically rich, economically sensitive, and covering a wider range of activities within our society than are currently considered. The dangers of letting these systems function without human oversight are illustrated by a growing list of real-life examples of algorithmic unfairness and errors causing social harm (see Angwin, Larson, Mattu & Kirchner, 2016; Datta, Tschantz & Datta, 2015). Leaving the human out-of-the-loop also poses questions of accountability (Bennett Moses & Chan, 2018; Diakopoulos, 2015). Accountability in this context refers to an obligation to explain or justify algorithmic decision-making, which is fundamental to mitigating negative social impacts or harms. Diakopoulos (2015) argued that human roles are already critical components in the creation of algorithms, during both the design and interpretation stages. Therefore, algorithmic accountability should actively reflect individual, group or institutional intent and the level of agency decision-makers have, when interpreting algorithmic outputs.

In practice domain experts are often unable to effectively use DSS predictions and simply choose to disregard them by returning to their old methods (even if less effective) (Brown, Chouldechova, Putnam-Hornstein, Tobin & Vaithianathan, 2019; Lee, Kim & Lizarondo,

Algorithmic decision support systems are widely applied in domains spanning a diverse range, from healthcare to journalism. To ensure that these systems are fair and accountable, it is essential that humans can maintain meaningful agency; understanding and overseeing algorithmic processes. Explainability is often seen as a promising mechanism for enabling this human-in-the-loop, however, current approaches are ineffective and can lead to various biases. We argue that explainability should be tailored to support naturalistic decision-making and sensemaking strategies employed by domain experts and novices. Using cognitive psychology and human factors literature principles, we map potential decision-making strategies dependant on expertise, risk and time dynamics and propose the conceptual Expertise, Risk and Time Explainability Framework, intended to be applied as explainability design guidelines. Finally, we present a worked example in journalism to illustrate the applicability of our framework in practice.CCS CONCEPTS • Human-centred computing → Human computer interaction (HCI); HCI Design and Evaluation Methods

* Corresponding author at: The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom.
E-mail addresses: a.simkute@sms.ed.ac.uk (A. Simkute), ewa.luger@ed.ac.uk (E. Luger), bronwyn.jones@ed.ac.uk (B. Jones), michael.evans@bbc.co.uk (M. Evans), rhia.jones@bbc.co.uk (R. Jones).

https://doi.org/10.1016/j.jrt.2021.100017
Received 29 June 2021; Received in revised form 14 October 2021; Accepted 17 November 2021
Available online 19 November 2021

2017). Spoon-feeding domain experts with algorithmic outputs, but depriving them of other important information, can leave them unable to understand, explain and justify their decisions and make use of algorithm-provided information (G. Klein, Moon & Hoffman, 2006b). The introduction of DSS can also disrupt domain experts' ability to apply their natural decision-making strategies, which can cause them to disregard algorithmic predictions (Lee et al., 2017), or demonstrate automation bias and overly trust them (Skitka, Mosier & Burdick, 1999). However, decision-makers are often held accountable for the outcomes even when they have little agency in the decision-making process (Wagner, 2019). One way to provide more agency is to use explainability techniques and inform decision-makers about the inner workings of the DSS and generation of the output. Explainability has received increased attention in recent years from researchers across various disciplines trying to find a way to make opaque AI and ML systems understandable to humans (Abdul, Vermeulen, Wang, Lim & Kankanhalli, 2018). Initially intended for ML experts, software engineers and data scientists (Arrieta et al., 2020), explainability approaches are now being used to support other stakeholders, such as users and domain experts (Tomsett, Braines, Harborne, Preece & Chakraborty, 2018). However, current explainability approaches lack usability and are not seen as effective by domain experts (Bhatt et al., 2020). There is also a risk that providing explanations could simply create a sense of unjustifiable trust and mislead decision-makers (Kaur et al., 2020).

We argue that for explainability to be effective, it is essential to understand how users interact with algorithms and what information is needed to support their decision-making strategies. To ensure that explanations can help decision-makers in maintaining meaningful agency, they should be tailored to support unique decision-making and sensemaking strategies of domain experts and novices. Few studies to our knowledge have attempted to explore human-algorithm interactions in a decision-making context (see De-Arteaga, Fogliato & Chouldechova, 2020; Green & Chen, 2019), and even fewer have examined factors influencing human decision-making and sensemaking strategies (see Simkute, Luger, Evans & Jones, 2020). Moreover, despite the many explainability techniques available, there are few design guidelines showing which method would be the most suitable in which situation, based on the decision maker's needs and contextual factors and considering differences in human reasoning or decision-making. There is also a lack of guidelines demonstrating how explainability could be integrated into existing applications that are used in real-world situations, for example, what to explain and how to display explanations in the interface as well as how to account for real-world constraints (Eiband et al., 2018). We suggest that a first step toward overcoming these issues should be building a solid understanding of naturally occurring human decision-making strategies and essential factors that influence them. To this end, we review decision-making literature, with a particular focus on decision strategies in naturalistic environments, expert decision-making, and decision-making in high-risk contexts. We outline several aspects that could help to predict which decision-making strategies will be followed depending on the level of risk, level of expertise, and time available. It is our intention that this knowledge might serve to inform which explainability heuristic would best support design strategy in any given situation. Based on these dynamics we have developed the Expertise, Risk and Time (ERT) explainability framework suitable for deployment and iterative development, with the long-term goal of supporting the development of effective design heuristics for explainable interface design, in a range of contexts. The contribution and purpose of the ERT explainability framework is to identify sensemaking strategies, cognitive biases, and attentional resources common to users of a predictive system and thereby assess the relevant explainability requirements. By offering three clear dynamics, we create a framework for designers seeking to scope out the explainability requirements in any given context. The scope of this article is not to develop detailed user interface (UI) designs, but to empower that design community, and we propose the future work needed to translate our insights and recommendations into UI design.

This paper makes three key contributions: 1) A systematic review of explainable AI research that highlights a need for more work on explainability in decision-making contexts in a much wider range of settings, including socio-technical domains that can be consider to have 'lower-stakes' and the potential for cognitive psychology and human factors literature to contribute useful insights for designing usable explanation interfaces; 2) The elaboration of a conceptual framework (ERT) to aid development of effective design heuristics for intelligible interface design, which could work as a tool to identify sensemaking strategies, cognitive biases, and attentional resources common to users of a predictive system and thereby assess explainability requirements;3) A worked example in the 'lower-stakes' context of journalism to demonstrate the usefulness of the ERT framework.

We begin by outlining the research methodology, reviewing motivations for, approaches to, and challenges of designing for explainability in algorithmic decision-making systems. We identify a need for further work in 'lower-stakes' decision-making contexts and draw insights from human factors and cognitive psychology literature to support effective explainability in decision-making contexts. The paper then describes the ERT framework, and elaborates a worked example in journalism of the framework's application. The closing sections discuss the usefulness, limitations, and future application of the framework.

2. Methodology

2.1. Literature review

The systematic literature review was structured thematically, focusing on three key themes, and related subthemes, that were identified gradually and searched in three stages. The first theme concentrated on current user-focused explainability approaches.

Stage 1: First, a small number of relevant and influential (most cited) papers in the field were identified as the initial bootstrapping stage in the structured literature search in an open manner. From these papers the author and index keywords were extracted and the most often occurring keywords were used as key search terms. The extracted terms were 'intelligible', 'explainable', 'interpretable', 'transparent' (with a condition of AI or ML also appearing in the abstract of the publication). The list of papers was manually filtered using set inclusion criteria and excluding papers that a) were presenting work in a different focus area; b) were only presenting/investigating algorithms, but not investigating intelligibility in a context where a human was present; c) did not consider a decision-maker as one of the stakeholders; d) were workshops or posters; d) were not in English language.

Stage 2: After conducting the first stage review, five reoccurring reasons for explainability from the user perspective were observed. Based on the keywords dominant in recognised relevant papers, the five themes were synthesised into the following key search terms: 'algorithmic accountability', 'algorithmic fairness', 'algorithmic transparency', 'human-in-the-loop' and 'trust' (with a condition of AI or ML or 'algorithmic systems' or 'decision support systems' also appearing in the abstract of the publication). The list of papers was manually filtered using a set inclusion criteria and excluding papers that a) were not investigating the decision-maker's perspective; b) presenting work in a different focus area; c) did not involve decision-making in either low- or high-stakes domain; d) were not full papers (posters or extended abstracts); e) were not in English language.

Stage 3: Following the first two stages of the review, the need for understanding the human agent was apparent; in particular, a more precise understanding of how they make sense of information provided and interact with algorithms in a decision-making context. Accordingly, during the third stage the focus was on how cognitive psychology science can support explainability. An initial review of most influential decision-making theories was conducted and cognitive psychology aspects relevant in the context of algorithm-supported decision-making

were selected and synthesised into four keywords. Key search terms 'human reasoning', 'mental models', 'expertise', 'sensemaking', combined with 'decision-making' were used in this stage. The list of papers was manually filtered using set inclusion criteria and excluding papers that were not investigating human psychology aspects in the decision-making context; or findings were not relevant/transferable in the algorithmic decision-making context.

2.2. Framework development and worked example

The characteristics recognised as influencing decision-making and sense-making strategies were extracted from the literature reviewed in the third review stage. Each characteristic then was weighted depending on a) the number of times it was mentioned or implied in the literature; b) strength of the characteristic (whether it outweighs the other characteristics); c) applicability (whether it is a general characteristic which can be determined in advance or an individual characteristic which must be measured at a personal level). The three highest weighting criteria were distilled: expertise, time-pressure, and task-related risks. Cognitive load, motivation, personality traits and uncertainty were the other considered characteristics, however, they did not meet the criteria, i.e., they were overpowered by the characteristics and were only measurable at the individual level. We distilled the decision-making strategies linked to each of the three characteristics, provided eight combinations of these strategies, for any given situation, and mapped existing explainability design approaches matching these combinations.

2.3. Scenario development

The final stage in our research methodology is the development of a worked case study in the context of journalism, to illustrate applicability of the ERT framework in practice. This involved creating two speculative scenarios containing narratives of how journalists could be reasonably expected to engage with two different decision-support systems and how a designer/design researcher could draw from the framework when observing these journalists in order to inform their explanation strategies. Scenario-based design methods anticipate and leverage scenarios of possible use at an early stage of system development and can be useful for understanding specific requirements that might arise when XAI systems are deployed into complex settings of use (Wolf, 2019). Using abductive reasoning, we developed scenarios in collaboration with an experienced journalist to test the potential utility of the framework and elicit aspects of the problem space DSSs open up in news production. We generated the scenarios by drawing on field data from an ongoing qualitative study of explainability and sensemaking of AI and algorithmic systems amongst journalists at a public service broadcaster. We supplemented this with secondary sources related to the case study domain, including (e.g. Diakopoulos, 2020, Gutierrez-Lopez et al., 2019). We also drew from the expertise of co-authors; one a professional journalist and a further two industry research and development practitioners embedded in a media organisation. Plausibility is a central criterion for validating scenarios as representational products and ensuring their heuristic effectiveness, i.e. that narratives are "derivable or can be arguably inferred or concluded from the initial conditions", making it reasonable to believe that they could happen, are trustworthy and are credible (Urueña, 2019, p19). Scenarios are selective and staged, "both grounded in current situations (as derived from field data) but also speculative in their articulations of how those situations may be changed because of a new system" (Wolf, 2019, p254). We tested the plausibility of the scenarios with an expert in the development and testing of newsroom technology and several journalists before incorporating feedback regarding context, probability and coherence.

3. Decision-making systems and human-in-the-loop

DSS are being increasingly used in 'high-stakes' domains, for tasks, such as predicting homelessness risk (Kube, Das & Fowler, 2019); screening for child maltreatment risk (Brown et al., 2019; Chouldechova, Benavides-Prado, Fialko & Vaithianathan, 2018); directing food donations (Lee et al., 2017); making bail decisions (Angwin et al., 2016); organising stop and search policing (Young, Katell & Krafft, 2019); navigating maintenance cases in aviation (Wanner, Heinrich, Janiesch & Zschech, 2020), or diagnosing illnesses (e.g., Ahmad, Eckert & Teredesai, 2018; Caruana et al., 2015). DSSs are also making inroads into areas that have less immediately obvious societal impact but which are socially significant nonetheless, such as media production and journalism. Besides the potential to improve the accuracy of decisions, algorithms are often linked to a range of social, ethical, and legal issues (Annany & Crawford, 2018), lack of accountability (Diakopoulos, 2015) and even unfairness towards certain groups or individuals (Barocas & Selbst, 2016). Thus, there is a growing demand for a human to maintain a meaningful agency and be able to oversee algorithmic processes. In this section, we will outline motivations and challenges related to maintaining a human-in-the-loop. We will also discuss the role of domain expertise in the algorithmic decision-making context.

3.1. Importance of a meaningful human agency

It is generally agreed that algorithms should rarely replace the human role completely but should instead be used to enhance people's decision-making (Citron & Pasquale, 2014; Green & Chen, 2020), and free up their valuable time for more thorough assessment of complex cases (Raghu et al., 2019), or creative work (Diakopoulos, 2019). Allowing algorithmic decision-making support systems to function without human oversight can lead to discrimination and perpetuation of biases. For example, ProPublica's analysis of the risk assessment system COMPAS demonstrated how it was discriminating against black defendants (Angwin et al., 2016). In another example, gender inequality was perpetuated by historically biased algorithms, when Google's job search engine showed men ads for jobs with higher pay than women (Datta et al., 2015).

Although so-called "low-stakes" domains are often overlooked and under-explored by researchers, consequences of algorithmic errors in these domains can also be costly and influence quality of life and well-being. 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, Eskens, van Drunen, Bastian & Möller, 2019). Moreover, use of audience analytics systems have been shown to shape normative considerations and editorial decision-making (Belair-Gagnon & Holton, 2018; Christin, 2020).

Without meaningful human input, algorithmic unfairness might remain unrecognised until targeted investigation is conducted (Angwin et al., 2016) and might lead to replication and even amplification of existing biases in society (Zhao, Chen, Wu, Chen & Liu, 2017). It also poses a question of who should be held accountable in cases where the algorithm misbehaves (Bennett Moses & Chan, 2018; Diakopoulos, 2015). It has been shown that decision-makers in areas such as autonomous driving or social media moderation, 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). Having a human-in-the-loop has been shown to be an effective way to reduce error rates in medicine (Raghu et al., 2019), and to increase fairness in recidivism risk predictions (Tan, Adebayo, Inkpen & Kamar, 2018). Affected members of society also express a preference for maintaining human agency – they believe that humans can better ensure consideration of any unusual or salient factors when making decisions (Brown et al., 2019). Moreover, people tend to trust a system more and show higher satisfaction with it if they know that the process is not completely automated (Lee et al., 2017). Human involvement can also ensure that decision-makers can explain and justify their decisions and use of

algorithmic outputs as a way to address accountability concerns (Wieringa, 2020).

3.2. Challenges of staying in the decision-making loop

However, decision-makers are often unable to evaluate the accuracy of the algorithms and make informed decisions (Green & Chen, 2019). In many instances they 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 basic understanding about the system they are using (Wagner, 2019; Young et al., 2019). For example, Young et al. (2019) observed government workers using algorithmic tools for surveillance and reported that they mostly lack knowledge about the system and drastically underestimated the complexity of it. Some of the employees were even unaware that algorithmic surveillance technologies rely on algorithmic or ML systems and assumed that computer vision should be an easily achievable task. Without a deeper understanding of a system's inner workings, decision-makers find it difficult to determine whether they should rely on algorithmic outputs in their decision-making (Yu et al., 2017). A decision-maker is considered out-of-the-loop if unable to identify irregularities and errors in the system, take corrective action when needed, and be held accountable in case the system misbehaves (Rahwan, 2018).

When decision-makers do not have means for building meaningful trust in the system, they are also more likely to be susceptible to various biases. For example, domain experts often succumb to automation bias when assessing reliability of the algorithmic predictions, i.e., they overrely on them. Airline pilots in the study by Skitka et al. (1999) showed a tendency to incorrectly follow predictive systems' advice. They made omission errors – failure to react to irregularities and faults if the automated system does not detect them; and commission errors – failure to properly assess the system's predictions and followed them despite the contradicting information from other sources (Skitka et al., 1999). Automation bias persisted even when pilots were accompanied by another pilot, were informed about potential risks of automation bias, or were trained to verify the automated recommendations and even when prompted for verification (Skitka, Mosier, Burdick & Rosenblatt, 2000) Automation appeared to simply reduce cognitive efforts put into decision-making by experienced pilots. Decision-makers might also express automation bias if they are not exposed to system's errors during the training process (Sauer, Chavaillaz & Wastell, 2016). In Sauer et al. (2016) study participants underwent training where they were exposed to either a fully reliable system or one of the three faulty systems: (a) faults detected and reported; b) faults detected and not reported; c) and faults not detected. When participants were asked to use the system after their training, they trained with undetected or unreported faults, were making more errors, and trusted system predictions more than their own knowledge failing to detect errors.

On the other hand, decision-makers can express distrust in the algorithms and automation in general, systematically disregarding predictions or refusing to rely on them (Veale, Van Kleek & Binns, 2018). This phenomenon is referred to as algorithmic aversion. It has been observed amongst lay users (Dietvorst, Simmons & Massey, 2015), as well as experts such as helicopter pilots (Veale et al., 2018). An ethnographic study by Whalen (1995) on emergency dispatchers using a new automated despatch decision-making system showed that they were reluctant to trust system outputs and checked them manually, even six months after the introduction of automation. More recent examples of algorithmic aversion by experts, was observed by Lee et al. (2017) who studied automation practices in food donation services and interviewed various stakeholders. One of the observations was that the community manager, making final food allocations based on the algorithmic analysis, continued using methods (her own heuristics and logic) adopted before introduction of automation to make allocation decisions for 1.5 years. Making decision-makers aware of any performance errors might also diminish their trust in the system and make them more likely to trust less accurate predictions made by humans instead (Dietvorst et al., 2015). For example, participants trusted their own less accurate speed dating predictions than the more accurate ones made by ML, if they observed any inaccuracies (Yin, Wortman Vaughan & Wallach, 2019).

3.3. Domain expertise in the algorithmic decision-making context

DSS are deployed in many settings in which human decision-makers are required to have a certain level of expertise. For example, air traffic controllers using algorithmic predictions to handle air traffic volumes must have sufficient experience of these specific tasks. Although in the real-world expertise seems to be lost to the expense of automation (Skitka et al., 1999), few studies have looked at how experts interact with the algorithms in the actual environment they are implemented in (e.g., Bussone, Stumpf and O'Sullivan (2015); De-Arteaga et al. (2020); Holzinger (2016)). Without providing means for experts to be involved in decision-making processes and have meaningful agency, there is a risk of losing the benefits of human expertise.

Although experts demonstrate incredible ability to quickly and intuitively spot irregularities in data or notice patterns that at first seem insignificant in naturalistic situations (Klein & Chase, 1998), they might be unable to apply their expertise when new factors, such as algorithmic support, are introduced (Sterman & Sweeney, 2004). Subsequently, decision-making in an algorithmically-supported context often suffers due to poor contextual fit of the system (Elwyn et al., 2013). Decision-makers might also struggle to apply their expertise and effectively use predictive systems due to disruption of their naturally occurring decision-making and sensemaking strategies (G. Klein et al., 2006b). It has been demonstrated that DSS change the nature of decision-making by users, and that of experts (De-Arteaga et al., 2020; Yang, Steinfeld & Zimmerman, 2019), leaving them feeling restrained by the static nature of the predictions (Yang et al., 2019) and unable to exercise their skills gained while working without algorithmic support (De-Arteaga et al., 2020).

In this way, changes in a particular setting can reduce effectiveness and disable use of the heuristics and other strategies learned with experience (Sterman & Sweeney, 2004). Even when the task maintains the same logical structure as before DSS are introduced, contextual changes might not allow existing skills to be transferred to the new environment (Sterman & Sweeney, 2004). Subsequently, both novices and experts, when introduced to that new environment, are likely to rely on their common sense or heuristics, thus underestimating other aspects and only searching for and accepting evidence that is consistent with their existing beliefs, leading to confirmation bias (Nickerson, 1998). Failure to appreciate the context in which decisions are normally made without the algorithmic support has been shown to be one of the main reasons why predictive systems fail in practice (Wagner, 2019). Poor contextual fit means that decision-makers might feel limited and resist relying on a system's predictions (Khairat, Marc, Crosby & Sanousi, 2018; Yang et al., 2019) or simply will not have means or time to make an informed decision (Wagner, 2019). Even predictive systems that focus on a specific task can fail if contextual factors are being ignored (Nickerson, 1998). For example, Veale et al. (2018) interviewed workers in public sector organisations and showed that how users interacted with algorithms, and whether they relied on them, depended on how well the system fit with their natural workflow and organisational context.

Introduction of predictive systems can also disrupt experts' ability to apply their natural decision-making strategies, especially if they are only shown the final output of already processed data (Klein et al., 2006a) Additionally, being spoon-fed interpretations can be frustrating and demotivating (Klein et al., 2006b). Recent studies demonstrated that experts felt that the outputs they received did not allow them to see a full

picture (Yang et al., 2019). Experts also expressed a preference for having access to raw features of the data systems so they could interpret it in the way they had been trained without algorithmic support (De-Arteaga et al., 2020). Studies showed that experts also sought ways to better understand the reasoning chain of the decision model (Bussone et al., 2015) and, being unable to apply their decision-making strategies, turned to their old methods (even if less effective), did not rely on the algorithmic predictions (Lee et al., 2017), or demonstrate automation bias.

In this section we showed that DSS are increasingly used in a wide range of domains where errors can have long-lasting societal consequences. Hence, meaningful human agency is important in preventing these errors and ensuring accountability in erroneous instances. However, decision-makers often fail to evaluate algorithmic predictions and are susceptible to automation bias and algorithmic aversion. Domain experts are also unable to use their unique decision-making abilities and expert knowledge when having to rely on DSS outputs.

4. Explainability in a decision-making context

Terminology in this domain, whilst important for understanding distinct but overlapping concepts, is still nascent and sometimes inconsistently defined. Explainability can be defined as the AI/ML model's ability to explain its inner workings and logic behind the output in human understandable terms (Doshi-Velez & Kim, 2017; Gilpin et al., 2018). Here, the term 'Explainability' is distinguished from a similar (and often interchangeably used) term - 'Transparency'. In the decision-making context the aim of explainability is to make relevant information available and understandable to the decision-maker (e.g., an experienced journalist), with the goal of supporting their sense-making process. Here we consider transparency as aiming to inform affected stakeholders (e.g., news media audiences) about algorithmic processes and methodologies (Diakopoulos & Koliska, 2017), giving them an opportunity to evaluate accuracy of the outputs (e.g., news stories) (Stark & Diakopoulos, 2016), and revealing the limitations of the given model (Coddington, 2015; Diakopoulos, 2014).

The need for explainability techniques has been growing dramatically (Fuji, Nakazawa & Yoshida, 2020). However, despite the immense research efforts, explainability approaches still lack usability and are ineffective when applied in a decision-making context (Abdul et al., 2018). In this section we will outline the ways explainability could support decision-makers, raise issues with current approaches, and discuss what could be done to make explainability more effective in the decision-making context.

4.1. How can explainability support decision-makers?

Aspirations for more accurate and powerful predictions have led to the widespread application of complex ML techniques, such as random forests and deep neural nets (Adadi & Berrada, 2018). Inherently interpretable, but arguably less accurate models are often traded for opaque, black box models that can be incomprehensible even to ML engineers and data scientists (Arrieta et al., 2020). Explaining a black box model (global explainability), or its output (local explainability) requires an explainability approach that would generate post-hoc explanations. Some cutting-edge approaches involve learning a simple local approximation of the underlying models around a particular data point (e.g., LIME, Ribeiro, Singh & Guestrin, 2016), and producing an additive feature importance score for single predictions (e.g., SHAP, Lundberg & Lee, 2017).

In general, explainability of AI/ML is intended to make the basis behind a system's reasoning in arriving at a prediction comprehensible to humans (Fuji et al., 2020). It should also reveal the strengths and weaknesses of a decision-making system and enable humans to predict future behaviours (Gunning & Aha, 2019).

Explainability could be an important tool allowing decision-makers to understand the logic behind ML predictions and enabling meaningful agency by providing relevant information (Cutillo et al., 2020; VanBerlo, Ross, Rivard & Booker, 2021). Explainability could help them establish understanding of how a system works, which is necessary for trusting a newly introduced ML/AI system (Brennen, 2020). Making model behaviour comprehensible to the decision-makers might reduce the cognitive load involved in performing the task (Fan et al., 2008), and help users overcome algorithmic aversion by providing a comfortable sense of understanding (Yeomans, Shah, Mullainathan & Kleinberg, 2019). Explainability could also give decision-makers a sense of control and in turn increase trust in the system (Dietvorst, Simmons & Massey, 2018; Kulesza, Burnett, Wong & Stumpf, 2015).

4.2. Issues with current explainability approaches

However, despite great efforts in explainability research, many of the proposed approaches lack usability when implemented in practice (Abdul et al., 2018). Lack of usability can result in the explainability attempts being ignored or misused by the stakeholders they are intended to help. For example, Bhatt et al. (2020) interviewed data scientists and other stakeholders across 30 organisations and revealed that explainability was mainly viewed as a tool for debugging the model used by ML engineers; users and decision-makers did not see explainability as a useful tool for them - they mainly thought of it as a tool designed for ML experts. Usability issues could be the result of a lack of appreciation of different stakeholders' explainability needs. Until recently, most research effort was focused on supporting ML experts and data scientists, often overlooking the needs of a wide range of other stakeholders seeking to comprehend the workings of opaque systems (Tomsett et al., 2018). Nowadays, researchers seem to agree that more attention needs to be paid to the needs of various stakeholders to ensure that explainability can be usable when applied in practice (Bhatt et al., 2020; Millecamp, Htun, Conati & Verbert, 2019; Rosenfeld & Richardson, 2019; Srinivasan & Chander, 2020).

Explainability in a decision-making context should also be used cautiously, with a clear goal of enhancing trust rather than just improving users' willingness to use the system and avoid algorithmic aversion (Liao, Gruen & Miller, 2020). Otherwise using explainability for building trust in the system and its predictions might create a sense of unjustifiable confidence (Yeomans et al., 2019) and result in automation bias (Kaur et al., 2020). Interviews with data scientists using popular explainability techniques revealed that these techniques were often misused and over-relied on in practice (Kaur et al., 2020). Some argue that explainability enables development of certain heuristics about the system, and users stop evaluating each individual decision or explanations (Bansal et al., 2021). Buçinca, Malaya and Gajos (2021) suggested using cognitive forcing intervention for people to engage with the AI-generated explanations more thoughtfully. Although this technique seemed daunting and made the system design less user-friendly, it was more effective in reducing overreliance compared to the standard explainability techniques (Buçinca et al., 2021). The amount of detail and information used in an explanation might also result in either automation bias or algorithmic aversion. Bussone et al. (2015) conducted a study with healthcare experts and reported that detailed and informative explanations indeed increased trust in the system and its outputs at the same time as increasing risk of overreliance. Informative and detailed explanation led medics to believe that the system used the best available medical knowledge, and similar reasoning processes as theirs. However, using less detailed explanations had an opposite effect and resulted in algorithmic aversion.

4.3. How can explainability be made useful for decision-makers relying on algorithms?

To overcome the usability challenges of explainability approaches, researchers have sought to find out in a more general way which

explanations different stakeholders would find most useful. Some looked for answers by analysing how people use explanations in real-life situations (De Graaf & Malle, 2017; Eiband et al., 2018; Garcia et al., 2018) and explored wider disciplines, such as psychology, philosophy, and cognitive sciences (Beaudouin et al., 2020; Hoffman, Miller, Mueller, Klein & Clancey, 2018; Miller, 2019; Srinivasan & Chander, 2020). One of the most in-depth works, linking multidisciplinary literature concerning explanations and explainability research, was by Miller (2019) who surveyed a vast number of empirical studies from the social sciences and presented core aspects of explanations. However, even knowing which explanations are generally preferred by people does not answer the question of how complex and detailed these explanations need to be in different instances.

There are multiple ways in which complexity of explanations can vary, making them either more or less comprehensible. For example, explanations can differ in a) size, and have a different number of lines and terms within the output clause; b) number of cognitive chunks, i.e., clauses of the output that may recur throughout the decision set, that can be implicitly or explicitly defined; or c) number of times the input conditions are repeated in the decision set (Lage et al., 2019). Complexity of the explanations can also vary depending on how sound (i.e., focused, and detailed explanations) and complete (i.e., explaining all the reasons) they are (Garcia et al., 2018). Moreover, explanations can be made interactive or conversational so that users can probe deeper until satisfactory understanding is achieved (Madumal, Miller, Sonenberg & Vetere, 2019; Weld & Bansal, 2018), or submit corrections and feedback (Smith-Renner et al., 2020). Tailoring the complexity and the content of explanations to the user or a context, might lead to more effective ways of explaining (Schaffer et al., 2015). Using tailored (case-specific) explanations instead of generic ones could also reduce the cognitive load of the task and help to avoid overwhelming users (Naiseh, Jiang, Ma & Ali, 2020). However, effective personalisation requires establishing what aspects should shape personalisation of explanations and how, and predicting the explainability needs of different stakeholders has proven to be complicated due to the complexity of the topic, multiple goals, and wide range of interested parties (Fu et al., 2020; Ras, van Gerven & Haselager, 2018).

We argue that to provide effective explainability, it is necessary to reflect on how decision-makers interact with DSS and what design choices could support their decision-making. To ensure that explainability can help domain experts and novices to maintain meaningful agency, explainability approaches should be tailored to enable their use of naturalistic decision-making and sensemaking strategies. Few studies to date have explored factors influencing human decision-making and sensemaking strategies in human-algorithm interactions (see Simkute et al., 2020). Morevoer, there is a lack of design guidelines, that would advice which explainability interface design approach would be the most suitable in which situation, based on the decision maker's needs and contextual factors. Exceptions include a set of usability guidelines by Amershi et al. (2019) and XAI Question Bank by Liao et al. (2020). However, the former, although relevant, is not specific to explainability and the latter is based on interviews with UX practitioners and designers, suggesting what users would want to know. Neither considers differences in human reasoning or decision-making. There is a need for guidelines that would demonstrate how explainability could be used to support decision-makers, for example, what to explain and how to display explanations in the interface as well as how to account for real-world constraints (Eiband et al., 2018).

Overall, explaining the logic behind DSS outputs to decision-makers, could provide them with more agency and help to build meaningful trust. However, explainability is rarely seen as a technique useful in a decision-making context. As we have seen, it can also lead to automation bias or algorithmic aversion. We argue that effective explainability should be tailored to support naturalistic strategies that domain experts and novices employ when processing information and making decisions.

5. Insights from human factors and cognitive psychology research

We suggest that a first step toward overcoming these issues should be building a solid understanding of naturally occurring human decision-making strategies and essential factors that influence them. To this end, we have conducted a structured, systematic review of cognitive psychology literature related to decision-making, with a particular focus on decision strategies in naturalistic environments, expert decision-making, and decision-making in high-risk contexts. We outline several aspects that could help to predict which decision-making strategies will be followed depending on the level of risk, level of expertise, and time available. It is our intention that this knowledge might serve to inform which intelligibility heuristic would best support design strategies in any given situation.

5.1. Decision-making in a high-risk context

Situations in which decisions are likely to have significant consequences, and/or can result in discrimination, damaged dignity, loss of credibility, or even loss of life or property, as well as situations in which the decision-maker faces high performance and social pressures can be described as high-risk situations (Orasanu, 1997). In high-risk contexts, decision-making strategies are dependant on aspects that can accelerate stress, in particular uncertainty of information (Orasanu, 1997), perceived lack of control (Breznitz, 1989) and time-pressure (Perlow, Okhuysen & Repenning, 2002).

In unfamiliar high-risk situations, where information is ambiguous or incomplete, decision-makers are believed to search for cues that would link the situation to any past experiences (Orasanu, 2005). The effects of stress are particularly high when these cues are unclear, cannot be recognised, assessed and matched (Orasanu, 1997). In these situations, decision-makers proceed with the cognitively demanding process of generating and matching multiple solutions, at the same time considering potential consequences (Orasanu, 1997). More precisely, decision-makers in high-risk contexts reduce uncertainty by matching the situation with similar past experiences, then generating and evaluating potential options serially one by one and, if time allows, mentally simulate potential scenarios (Lipshitz & Strauss, 1997). This strategy places a high load on the decision maker's working memory, as multiple goals and strategies have to be held in working memory while the constraints are retrieved and evaluated (Orasanu, 1997).

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. Decision-makers generally show overconfidence in their decision-making and forecasting abilities (Tversky & Kahneman, 1974) and make errors by overestimating their impact on the outcome (Langer & Abelson, 1983). Thus, they are susceptible to the illusion that their predictions are valid and tend to overcommit to their choices (Einhorn & Hogarth, 1978). On the other hand, in high-risk situations, the decision-maker is more likely to be motivated to employ thorough and analytical strategies of decision-making. A highly motivated decision-maker is more likely to challenge heuristics and slow down the process of decision making (Svenson, 1979). They are more likely to spend more time evaluating all available options and avoiding making fast and intuitive decisions (Svenson, 1979).

Decision-makers show better performance and ability to evaluate available information in high-risk situations when they can use a certain rule-base protocol or a checklist in order to reduce ambiguity (Orasanu, 1997). Research in a medical setting showed that effects of uncertainty could be minimised by helping decision-makers to match the situations to a certain rule-based protocol (Dobrow, Goel, Lemieux-Charles & Black, 2006). According to study participants, support tools such as decision principles and evidence hierarchies were essential in revealing important modifiers that they were able to recognise and use for

decisioning (Dobrow et al., 2006). Decisions that follow a set of rules or a specific checklist are the least 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, 1997). In this way cognitive load is lowered, and an analytical decision-making process becomes possible.

5.2. Decision-making under time-pressure

Decision-making strategies are highly affected by the time available to process the information and come up with the best solution. A strong sense of urgency has been shown to influence the pace of decision-making (Perlow et al., 2002). In general, time-pressure is believed to have a negative effect on decisioning effectiveness, and general performance (Oliva & Sterman, 2001). Not pressured by time constraints, the decision-maker is likely to apply complex decision strategies in order to find the most logically suitable solution (Svenson, 1979). This way the decision is made based on thorough and detailed analysis of options, and is most likely to be the best available, with the highest probability of success. On the other hand, when decisions have to 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 conducting a full search of relevant information (Svenson, 1979). Decision-makers pressured by time are more likely to process information serially by generating and evaluating one option at the time until one that is reasonably fitting is accepted (Klein, Calderwood & Clinton-Cirocco, 2010). When decision-making is constrained by time-pressure, the amount of information presented can influence the effectiveness of the decision made and help to regulate cognitive demands put on the decision-maker. For example, providing multiple alternatives does not help to reach more valid or reasonable decisions as, due to sequential processing, these can be cut short as soon as an acceptable option is met, leaving the rest of the alternatives unconsidered (Klein & Chase, 1998). Similarly, providing too much information and/or additional resources can lead to ineffectiveness in decision-making (Omodei, Wearing, McLennan, Elliott & Clancy, 2005). Firstly, there is a limit of 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 all available information resources, as well as providing a huge amount of information, can be counterproductive (Omodei et al., 2005). This is mainly believed to be due to the overutilising bias – the bias to exploit all resources, happening outside conscious awareness (Reason, 1990). Decision-makers are intended to believe that they can effectively manage information and resources available, however, they do not appreciate the limitations of their ability to regulate the related cognitive workload (Omodei et al., 2005).

Secondly, in time-pressured contexts, the decision-maker might feel urgency to use the available resources, whether that would be information gathering, opportunity for action, or communication input (Omodei et al., 2005). The overutilisation of resources does not stop even when the cognitive system of a decision-maker is overloaded and damages their ability to make effective decisions. Thus, access to multiple information sources can indeed be disadvantageous (Seagull, Wickens & Loeb, 2001). Besides overutilising bias, the tendency to overutilise information also comes from other general biases. For example, under time-pressure, commission errors are preferred over omission errors, meaning that decision-makers prefer to make mistakes when proceeding with action, rather than due to the delay and inaction (Kerr, MacCoun & Kramer, 1996). Acting instead of waiting, even if ineffectively, also brings an illusion of control over the task, the sense of achieving some results (Schmitt & Klein, 1999), and a sense of greater self-competence via activity (Dörner, 1990). Due to the illusion of control (Duhaime & Schwenk, 1985) decision-makers make errors by overestimating their abilities and their impact on the outcome (Langer & Abelson, 1983). They may assume that through additional effort they can make their strategy succeed should problems arise (Langer & Abelson, 1983). 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 (Camerer, Johnson, Rymon & Sen, 1993).

5.3. Expertise and decision-making

Experts have been shown to engage in intuitive decision making rather than detailed analysis of all the options made available to them (Klein & Chase, 1998). With little conscious consideration, experts follow the route that intuitively seems most suitable and rarely consider more than one option (Klein & Chase, 1998). Intuitive decision-making does not mean that experts make important decisions carelessly. Indeed, there is evidence that skilled decision-makers often do better when they trust their intuitions rather than when they engage in detailed analysis (Klein, 2003). For example, Benner, Tanner and Chesla (1992) showed that experienced nurses who assessed the situation using their existing expertise, with little analytical efforts, were able to identify unusual and important information that otherwise might have been ignored. On the other hand, novice nurses who had to rely on a protocol-based checklist, were only able to diagnose, but not anticipate and prevent illness. They were also able to come to the solutions faster by generating fewer options that need consideration, whereas novices would have to produce a number of options and conduct analytical comparison of them (Cesna & Mosier, 2005). Experts are also particularly sensitive to the context and thus are better at noticing features of situations that could have potential implications (Klein et al., 2010). According to the recognition-primed decision (RPD) model, experts rely on pattern recognition when making decisions (Klein, 2003). They tend to quickly and subconsciously recognise patterns and cues in situations (or a data set) and intuitively link them to other cues that they expect to appear next (Schmitt & Klein, 1999). According to this model, high-expertise decision-makers rely on their past experiences in order to recognise cue patterns that allow them to understand and evaluate the problem or information. The cue recognition triggers retrieval of a response, which is drawn from the similar past experienced with matching cue patterns (Orasanu, 2005). Seeing the noise of the data, not only the main trends, guides expert's intuition and triggers the pattern recognition, which in turn allows them to know which cues to monitor and which are important or doubtful (Ross, Shafer & Klein, 2006). Experts under time-pressure are also more likely to engage into RPD types of decision-making (Klein & Chase, 1998). However, the RPD decisions are unlikely when decision requires justification, as intuitive decisioning is hard to articulate (Orasanu, 2005).

When a decision is particularly important, and experts have to test a hypothesis in an analytic way, they use schema; a pack of domain-specific knowledge they possess (Coderre, Mandin, Harasym & Fick, 2003). For example, experienced doctors have been shown to use knowledge templates built on previous experiences when making diagnostic decisions (Sibbald, de Bruin & van Merrienboer, 2013). When use of experience is impossible and experts have to search for information in order to reduce uncertainty, they benefit from being allowed to freely explore the information (Schmitt & Klein, 1999). Their search strategy of relevant information depends on material encountered, rather than following neat or orderly methods (Camerer et al., 1993). Experts decide what further information they need during the process of decision-making, not in advance. The static nature of algorithmic predictions and lack of freedom in exploring raw data has also been criticised by experts in recent studies (De-Arteaga et al., 2020; Yang et al., 2019). When the expert is unable to freely explore data, or only receives an input without knowing the inner workings of the system, their analysis suffers (G. Klein et al., 2006a). Moreover, experts have been shown to prefer to be involved and be able to actively question data when it is inconsistent with their intuitions. Being able to participate in active search, adaptation and mental model building processes allow

experts to better exercise their expertise and maintain their motivation (G. Klein et al., 2006a).

Expertise can also lead to types of error that should be considered in selecting explainability design heuristics. For example, a high level of expertise can lead to "illusion of validity", i.e., an unjustified sense of confidence and hence failure of evaluating different possibilities (Kahnemen & Klein, 2009). Experts make errors when over-relying on shallow processing and do not question their intuition (Eva & Cunnington, 2006). A chance to formulate and test different scenarios and see various ways to account for the same data can help experts to overcome their bias towards their intuition (Klein, 2003). Whereas simply providing digested data outputs or explanations of them, could lead to biases towards the most salient or intuitively most likely decision (Kahnemen & Klein, 2009). The way in which information is presented can also either help or further disrupt the expertise in decision-making. Providing more information than is necessary can damage an expert's performance (G. Klein et al., 2006a). Too much information (especially under time-pressure) could lead to overconfidence in experts (G. Klein et al., 2006a). and place an unreasonable cognitive load on a decision-maker (Klein et al., 2010). When rapid decisions are needed, it is more effective to provide information sequentially, as only one option is being considered at the time. However, RPD should also be encouraged in cases when satisfaction criteria for the solution is enough, and optimising is not necessary.

Overall, understanding naturalistic decision-making strategies could help to improve explainability potential. Decisions made in high-risk contexts are more cognitively demanding and require reducing uncertainty by matching an unclear situation with similar past experiences. In these situations, decision-makers perform better if they can follow a rule-based protocol as they are more motivated to slow down their decisioning and use analytical thinking. Time-pressure can impair decisioning and pressure the decision-maker to overutilise all available information. Limiting the amount of information can help to overcome this overutilisation of resources. Lastly, experts are more likely to make decisions intuitively, based on the patterns and salient features they recognise in the context. Their decision-making can be aided by providing noisy data, allowing them to apply their random decision search strategies and actively engage with data.

6. Contextual ERT framework for explainability

Despite the attention DSS have received in recent years, little has been done to grasp the psychology behind the human-algorithm interactions in a decision-making context. Although explainability researchers have made attempts to consult psychology research literature in search for more effective explanations, they only offer very broad understanding of the types of explanations people generally find more comprehensible (Miller, 2019). Even when proposing more 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 goals (Hind et al., 2019) rather than naturalistic ways of decisioning. We recognise higher-level decision-making and sensemaking strategies that are domain-agnostic and argue that to enable use of expertise and effective decision-making, these strategies need to be supported.

The review of cognitive psychology and human factors research literature concerning expertise and decision-making and sensemaking strategies revealed several dynamics that are particularly important in shaping how people make decisions. More precisely, level of expertise, level of contextual risk, and time constraints were recognised as factors influencing sensemaking strategies, cognitive biases and attentional resources in a decision-making context. Based on the reviewed literature we mapped decision-making strategies that are likely to be taken 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 a high and low levels of expertise and risk. Each section is then moderated by the level of time-pressure in each context, dividing decision-making space into eight segments. Each segment represents different information processing strategies, cognitive biases and attentional resources in a given combination of dynamics (Table 1). The framework is intended to help match these aspects with suitable design approaches and characteristics of explanations.

The ERT explainability framework is suitable for deployment and iterative development, with the long-term goal of supporting the development of effective design heuristics for intelligible interface design, in a range of contexts. By offering three clear dynamics, we create a framework for designers seeking to scope out the explainability requirements in any given context. The following section offers a

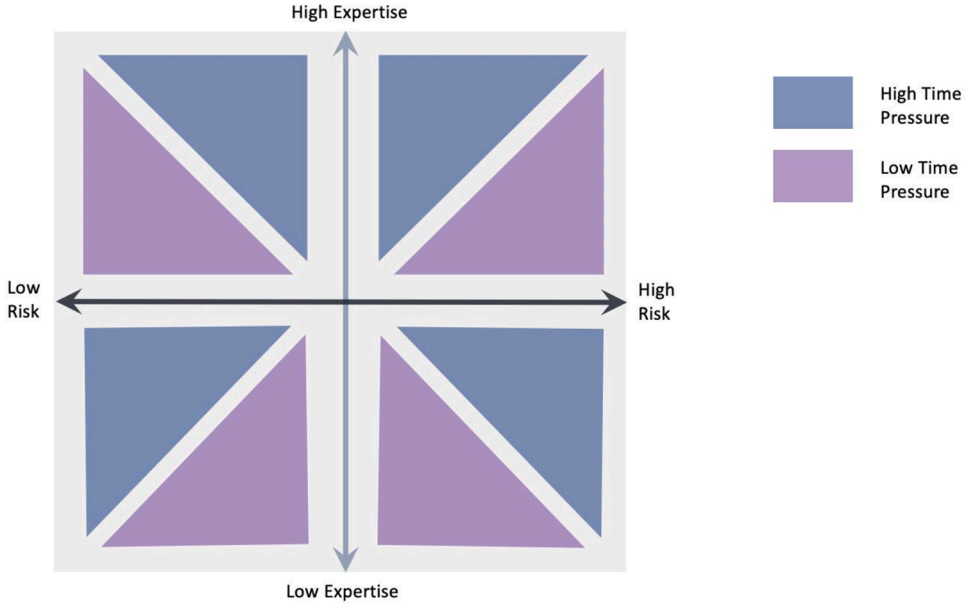

Fig. 1. Each of the eight segments represents different information processing strategies, cognitive biases and attentional resources that shape an overall decision-making process in a given combination of dynamics.

Table 1
Table linking decision-making strategies with design suggestions in the various combinations of ERT dynamics.

		High risk & high time pressure	High risk & low time pressure	Low risk & high time pressure	Low risk & low time pressure
Expert	Decision-making strategies and biases	Intuitive, fast, highly motivated, disruptedRely on pattern recognition and knowledge templates built on past experiencesUnder high uncertainty,generate and sequentially evaluate (mentally simulate) potential scenarios one at a time, until one that is reasonably fitting is acceptedSearch for salient features that could have implicationsOverconfidence biasIllusion of validity	Slow, analytical, highly motivatedRely on pattern recognition and knowledge templates built on past experiencesApply thorough and detailed evaluation of options, looking for the most logically suitable solutionSearch for salient features that could have implicationsOverconfidence biasIllusion of validity	Intuitive, fast, disrupted, low motivationQuickly and subconsciously recognise patterns and cues in situations (or a data set) and intuitively link them to other cues they expect to appear nextRarely consider more than one optionUnder uncertainty, generate few potential options that need considerationOverconfidence biasIllusion of validity	Intuitive, low motivationQuickly and subconsciously recognise patterns and cues in situations (or a data set) and intuitively link them to other cues they expect to appear nextRarely consider more than one option. Under uncertainty, generate potential options that need considerationOverconfidence biasIllusion of validity
	Design Approach	Present information sequentiallyUse rule-based protocols, checklists, decision principles, evidence hierarchies	Provide flexible ways to explore informationShow noise in dataUse rule-based protocols, checklists, decision principles, evidence hierarchiesProvide interactive ways of questioning explanations and providing feedback	Present information sequentiallyEmphasise the most important information	Limit the amount of informationEmphasise the most important information
Non-expert	Decision-making strategies and biases	Analytical, disrupted, highly motivatedApply thorough and analytical analysis of all possible optionsInability to effectively manage information and regulate cognitive workloadOverutilisation biasUnable to notice patterns in data and recognise salient features	Slow, analytical, highly motivatedProduce multiple possible options and conduct analytical comparison of them, looking for one with highest possibility of successUnable to notice patterns in data and recognise salient features	Analytical, disrupted, low motivationAnalyse all possible optionsInability to effectively manage information and regulate cognitive workloadOverutilisation biasUnable to notice patterns in data and recognise salient features	Slow, analytical, low motivationAnalyse all possible optionsUnable to notice patterns in data and recognise salient features
	Design Approach	Present only the most important aspects of information in a structured wayUse rule-based protocols, checklists, decision principles, evidence hierarchies	Present all aspects at the same time but guide the decision-maker towards the most important features or patterns.Use rule-based protocols, checklists, decision principles, evidence hierarchies	Present only the most important aspects of information in a structured way	Information is available to be explored and feedback is possible

description of methods used in the development of the framework and detailed explanation of the dynamic aspects, followed by an applied example with indicative design considerations.

6.1. The three dynamics

The ERT explainability framework divides the decision-making space into eight segments (Fig. 1). Each segment refers to the combination of (1) the extent to which decision-makers rely on their level of expertise in the particular task, (i.e., prior experience in making decisions without any algorithmic support), (2) the risk environment in which decisions are made (i.e., the cost of error) and (3) time constraints (i.e., cost of delay; given time to complete the task).

Dynamic 1: Expert or a novice

Firstly, explainability design strategies should be adjusted depending on whether the system will be used by an expert or a less experienced person. An expert in this context is "a trained professional with experience in some special domain" (Webster's New, 1968), whose expertise is a result of "a rich instrumental experience in the world and extensive and deliberate practice and feedback" (Hoffman, Shadbolt, Burton & Klein, 1995). Whether a person is an expert can be established by their reported familiarity with the task (Schaffer, O'Donovan, Michaelis, Raglin & Höllerer, 2019) or the type and quantity of experience in the actual domain or project (e.g., Ericsson & Smith, 1991; Hoffman et al., 1995). In our framework we only separate two levels of expertise: high level, i.e., decision-maker is an expert, and low level, i.e., decision-maker is familiar with the domain and/or task but is not an expert yet. Although there are varying levels of knowledge in every domain, research shows that only after reaching the level of being an expert, decision-makers begin to employ significantly different decision-making strategies than the people in other knowledge categories (Dreyfus & Dreyfus, 1986). Progression to expertise happens when a person advances from a superficial and literal understanding of problems to an articulated, conceptual, and principled understanding (Hoffman, 1998).

A system interface designed for the needs of a less experienced person, but applied for experts, might be demotivating for an expert. It could also result in either automation bias or algorithmic aversion, encouraging overconfidence and overreliance on heuristics. Failure to consider decision-making strategies applied by experts could lead to experts' inability to exercise their unique skills and recognise particularities that they learned to notice over years of accumulated experience. However, a system interface designed for an expert could overwhelm any novice decision-makers that use it, leading to errors and not allowing development of expertise.

An interface designed for explainability that is tailored for experts' needs should allow them to explore data more freely because when the expert is unable to freely explore data, or only receives an input without knowing the inner workings of the system, their analysis suffers (G. Klein et al., 2006a). For example, expert weather forecasters refused to use algorithmic metrics provided by a computer system, as these were "too smooth" and only revealed main trends of data but not the noise in data (Stuart, Schultz & Klein, 2007), similarly medical experts resist the algorithmic support because the static data would not allow them to see the full picture of the patients (Yang et al., 2019). An explainability interface should uncover these cues and patterns and lead 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. In particular, medical experts showed a preference for seeing which factors were most influential towards the predictions, as this would allow them to see which factors were modifiable and in turn plan future actions and interventions (Yang et al., 2019). De-Arteaga et al. (2020) pointed out that it is important to provide access to raw features of the data system and values of the features weighting towards the prediction. Expert users might be reluctant to use DSS if they are not able to 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, Westin & Borst, 2014). Experts' preference for flexibility in information search strategies could also be fulfilled by providing a level of interactivity (Weld & Bansal, 2018) or introducing refinement tools, that could guide the search process (Cai et al., 2019) and allowing decision-makers to correct errors (Kulesza et al., 2015).

Dynamic 2: Level of risk

Secondly, explainability design heuristics should be tailored depending on the level of risk assigned to the task. The complex nature of high-risk situations makes it difficult to draw predefined domain-agnostic risk criteria. Broadly, this paper considers the high-risk context as involving decisions that can have major consequences and high cost of error. The level of risk might depend on aspects such as the *extent* of the consequences (the chain of people affected by the decision e.g., journalist decision-maker, news organisation, audience, society), the temporality or permanency of the consequences (How long will these consequences be present? Is an option to undo the decision or adjust available?), and 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 design, could help to facilitate decision-making (Rundmo, 2001). Individual perception of risk has been shown to be unreliable (Marek, Tangenes & Hellesoy, 1985), depending on subjective feelings about technology (Slovic, Finucane, Peters & MacGregor, 2004), and highly varying at an individual level, even between individuals with directly comparable grades of expertise (Brehmer, 1992; William & Noyes, 2007).

Designing an interface capable of communicating the risk of the decision that is being made could calibrate the decision-maker's perception of risk. For example, an explainability interface could include pop-up alerts containing information regarding the risk or by using words (e.g., critical) that would indicate the level of severity of the situation (Long et al., 2018). Experts show higher consensus for linguistic risk representations than for numeric ones, thus appropriately selected words can more directly convey the risk (Atoyan, Robert & Duquet, 2008). Risk could also be visualised and indicated by using colours or symbols (Rayo & Moffatt-Bruce, 2015).

Using design approaches to inform decision-makers about the risk of the decision could also reduce the decision-maker's cognitive load. Experts' evaluation of risk requires situational awareness, and making additional diagnostic decisions (Kaempf, Klein, Thordsen & Wolf, 1996). Experts understand current situations by matching observed features with their previously learned interpretations of cues and patterns, and by mentally simulating a story that would explain how any given situation has occurred (Lipshitz & Cohen, 2005). Ordering and visualising features (cues) by their weight toward the output, could help to facilitate building of a coherent story that explains the available evidence (Liebhaber & Feher, 2002) Contextualising explanations by using visual examples could also ease feature matching and story generation/evaluation and thus reduce cognitive efforts of experts (Kaempf et al., 1996). Finally, using dynamic annotated visualisations instead of simple text-based aids, could help to effectively promote comprehension and present the risks of all the available decision options, without disrupting decision-makers' workflow (Rayo and Moffatt-Bruce, 2015).

Decisions made in high-risk context, with high cost of error require a slower and more analytical approach. In this case the decision-maker should be encouraged to gradually inspect information provided (Klein et al., 2010). However, the risk of overutilisation bias should be considered (the bias to exploit all resources), happening outside conscious awareness (Reason, 1990). Decision-makers tend to believe that they can effectively manage information and resources available, however, they often do not appreciate the limitations of their ability to regulate the related cognitive workload (Omodei et al., 2005). Thus, one

of the explainability interface design goals should be to slow down the decision-making process to avoid heuristics-driven decisions. This can be achieved by using interactive interfaces (Cheng et al., 2019), requiring acknowledgement of the explanation (Atoyan et al., 2008), or including an extra action needed to access the explanation (Rundo, Pirrone, Vitabile, Sala & Gambino, 2020). Interactive interfaces could also help experts to simulate mental scenarios when evaluating available information and determining the best course of action (Cheng et al., 2019).

In high-risk situations, decision-makers can also be supported by allowing them to identify effective options more easily. These options can be detected by comparing the results, situations to provided prototypes (Klein et al., 2010), rule-book protocols or checklists (Dobrow et al., 2006; Ross et al., 2006), easing the recognition of atypical situations that need action and amendments (Klein et al., 2010). This could help to reduce stress and related cognitive strain Orasanu (1997).

Level of risk also affects the motivation of the decision-maker. In high-risk situations, the decision-maker is highly motivated, so more information could be provided and critically evaluated (Svenson, 1979). Moreover, higher motivations also mean that decision-makers are less susceptible to biases and are more likely to critically challenge provided information and in case of high expertise – challenge their own intuition. On the other hand, low-risk contexts might lead to low motivation and shallow processing without critically challenging the information (Kahneman & Klein, 2009). This could result in automation bias. High motivation can be maintained by introducing a level of control to the users, e.g., ability to correct errors (Kulesza et al., 2015) or make modifications (Dietvorst et al., 2018).

Dynamic 3: Time-pressure

Both aspects are moderated by time pressure, as this affects what strategies a decision-maker will be able to employ. Under severe time constraints, when a slow and analytical approach is not possible, decision-makers are particularly susceptible to various biases. Not addressing this factor could lead to errors when decision-makers fail to judge their ability to accurately fill in the gaps in information and make assumptions.

Time limitations can be addressed by moderating the amount of information shown to the decision-maker. In high-risk and high time-pressure situations the design goals for an explainability interface should be to reduce clutter by limiting the number of alternatives shown at a time, whilst allowing access to detailed information once a particular alternative is selected. Too much information may trigger a utilization of simplifying heuristics, where the user can fail to focus the attention on important information. When rapid decisions are needed, it is more effective to provide information sequentially, as only one option is being considered at a time (Klein et al., 2010). In this way, faster and more intuitive decision-making strategies can be supported, and cognitive load reduced.

Visualising information can also reduce cognitive load, especially when a large amount of information needs to be processed by the user in a short period of time. It has been shown that experts in time-constrained situations use mental imagery when considering information, recognise cues and visually "paint" possible ways of implementation and potential outcomes (Klein et al., 2010). For example, visualising information in the form of a diagram can reduce the strain put on a working memory when processing and speed up the process of comprehension. When information is presented in a visual form, the user does not have to hold and later recover all the points of information in their working memory (Johnson-Laird, Legrenzi, Girotto & Legrenzi, 2000). Salient visualisations can be used for guiding less experienced decision-makers to the critical information (Eick & Wills, 1995) whilst, in contrast, experts would benefit from being able to freely explore the algorithm through interactive visualisations., i.e., by changing the attribute values and observing how the algorithmic decision changes accordingly (Cheng et al., 2019).

Having explained the three dynamics affecting decision-making, and suggested aspects of design that might be brought to bear in support of them, we use the next section to apply the framework to the practical expert craft of journalism.

6.2. Proposed design goals and examples of design strategies

Here we propose a list of design goals tailored to the decision-making strategies used under various combinations of expertise and risk dynamics, moderated by the time-pressure dynamic. Each design goal is followed by the potential design strategies that could be used in designing explainability interfaces. Examples with no indication of the time dynamic are applicable for both types of contexts.

High level of expertise and high-risk

- Calibrate the perception of risk: use pop-up alerts and/or linguistic indications informing about the risk, using colour and symbols to visualise the level of risk (Long et al., 2018)
- Facilitate pattern recognition, mental simulation/evaluation of alternative scenarios, reduce cognitive load: embed explanation information in the context (Kaempf et al., 1996), enable use of rule-based protocols (Dobrow et al., 2006)
- Support ability to expand information and see 'noise' in data: allow exploration of multiple variants within categories using refinement tools and clustering techniques (Cai et al., 2019)
- Support slow analytical decision-making in **low time-pressure**: allow interactive manipulation of attribute values and observe how the output changes accordingly (Speier & Morris, 2003), provide an ability to compare and contrast hypothesis/ features/ categories (Cai et al., 2019)
- Support serial information processing in **high time-pressure** situations: provide an option to view a single information point at a time, allow an easy transition to the next option (Klein et al., 2010)
- Support the ability to quickly recognise critical information in **high time-pressure** situations: use visualisations indicating critical information (Eick & Wills, 1995), indicate the predictive markers that are highly valuable (Long et al., 2018)
- Reduce information clutter in **high time-pressure** situations: limit the amount shown on the interface, instead allow to expand each data point by e.g., hovering the cursor over it (Cheng et al., 2019)
- Effectively promote understandability of weights of all available options, without disrupting the workflow: use dynamic annotated visualisations (Rayo and Moffatt-Bruce, 2015).
- Slow down decision-making process and prevent use of heuristics: use interactive interface (Cheng et al., 2019) and allow decision-maker to actively question the data, require explicit acknowledgement of the explanation (Atoyan et al., 2008), or require deliberate action to access explanation, instead of it being available but unremarkable (Rundo et al., 2020)
- Support use of flexible information search strategies in **low time-pressure** situations: provide ways to flexibly explore available information through interactive interface and adjustable inputs, allowing detailed information to be obtained by selecting data point/ feature (Cheng et al., 2019), or giving explicit control of which hypothesis/ features/ categories to compare and contrast (Cai et al., 2019)

High level of expertise and low-risk

- Facilitate pattern recognition, mental simulation/evaluation of alternative scenarios, reduce cognitive load: embed explanation information in the context (Kaempf et al., 1996), enable use of rule-based protocols (Dobrow et al., 2006)
- Allow flexible information exploration in **low time-pressure** situations: use refinement tools to guide the search mechanisms (Cai

et al., 2019), allow exploration of features in cases of disagreement or uncertainty
- Make explanation part of the workflow: present explanations in a seamless way, avoid interrupting features, use visual aspects such as colour-coding and symbols consistently (Yang et al., 2019)
- Reduce clutter and cognitive load in **high time-pressure** situations: limit the amount of information shown on the interface, instead allow to expand each data point by e.g., hovering the cursor over it (Cheng et al., 2019)
- Support serial information processing in **high time-pressure** situations: provide an option to view a single information point at a time, allow an easy transition to the next option (Klein et al., 2010)

Low level of expertise and high-risk

- Calibrate the perception of risk: use pop-up alerts and/or linguistic indications informing about the risk, using colour and symbols to visualise the level of risk (Long et al., 2018)
- Facilitate guided exploration of information in **high time-pressure** situations: use visualisations indicating critical information and showing the path of information exploration (Eick & Wills, 1995).
- Reduce cognitive load: enable use of checklists (Lipshitz & Cohen, 2005; Orasanu, 1997)
- Support the consideration of time available in **high time-pressure** situations: adjust the length of the explanation depending on how much time is available (Lipshitz & Cohen, 2005)
- Reduce clutter in **high time-pressure** situations: provide less detailed explanations, but clearly highlight the information that is critical also illustrating its criticality (Long et al., 2018), use bar charts to illustrate the breakdown of the decision and weight of different attributes towards the final output, group attributes by the colour (Cheng et al., 2019)
- Support analytical evaluation of all the available options in **low time-pressure** situations: use detailed descriptions of the attributes and features (Cheng et al., 2019), visualisation techniques to make it easier to compare different options, use dynamic visualisation and colour coding to illustrate each feature's weight towards the output, use interactive refinement tools to show changes in the distributions after updates (Cai et al., 2019)

Low level of expertise and low-risk

- Support structured information search strategies and facilitate building of coherent story in **high time-pressure** situations: order the features from the most to the least important (Liebhaber & Feher, 2002)
- Facilitate guided exploration of information in **high time-pressure** situations: use visualisations indicating critical information and showing the path of information exploration (Eick & Wills, 1995).
- Support analytical evaluation of all the available options in **low time-pressure** situations: use detailed descriptions of the attributes and features (Cheng et al., 2019), apply visualisation techniques to make it easier to compare different options, use dynamic visualisation and colour coding to illustrate each feature's weight towards the output, use refinement tool to show changes in the distributions after updates (Cai et al., 2019)
- Facilitate learning: allow interactive questioning of the output and provide feedback option

7. Worked example of the ERT framework: Journalism

Algorithmic and ML-driven DSS are gaining traction in journalism, where journalists increasingly rely on them for the gathering, production and distribution of news (Beckett, 2019; Diakopoulos, 2019; Marconi, 2020). Recommender engines drive new forms of audience personalisation and engagement (Helberger, 2019), 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, 2020), deeper analysis (Stray, 2019) and greater creativity (Maiden et al., 2018) but also risk disrupting long-established ways of working. Newsworkers consider 'journalistic intuition' and 'gut instinct' to be 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 algorithmic intervention, for instance toward placing more value on analytics than professional intuition (Hanusch, 2017). Most journalists have little understanding of how these systems work and limited ability to critically assess automated outputs or their suitability in context (Jones forthcoming). This knowledge and communication gap risks leading to journalistic malpractice (Hansen, Roca-Sales, Keegan & King, 2017) and undermining public confidence in ethical and responsible journalism. If this is to be avoided as DSS 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 the journalism context 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 in a timely manner 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 immediate and extreme (e.g., life and death decisions in defence, medicine etc.), the risk of opaque DSS in news production is one of aggregated errors, unrecognised biases and cumulative oversights. This can lead to inaccuracies and the inability to sufficiently account for and justify editorial decisions, which in turn can harm news organisations' reputation and undermine the legitimacy of journalism in society. Diakopoulos notes that journalists express a "need for ongoing scepticism and verification of outputs of data-mining" of the type used in DSS (2019, p.89) and points to the challenge of evaluating reliability of results. Knowledge claims in journalism are subject to varying criteria for adequate justification, which complicates the task of communicating the role of DSS in decision-making to both editors and the audience. Diakopoulos highlights the importance of building smart interfaces to support journalists, including by designing suitable signals to highlight relevance and other indicators of newsworthiness, as well as engender appropriate trust by reflecting uncertainty. This can include "multimodal and interactive interfaces", "summary sentences of text" providing explanation, or "visual evidence and context" (Diakopoulos, 2019, p82). The ERT framework can contribute to responding to these challenges. In the following section, we employ Wolf's concept of explainability scenarios (2019) - a scenario-based design approach to provide narrative descriptions of envisaged usage episodes to guide the development of a system (Rosson & Carroll, 2003). We use scenarios here to elucidate an example of when journalists interacting with DSS could require explanation alongside how designers might deploy the ERT framework to help design such explanations.

7.1. Scenario 1: Image suggestion decision-support tool

Leila is part of a team designing an explanation interface for an image suggestion system that uses ML 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 be able to justify to their editor why they chose the images and to 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 prototype of the system, 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 meeting 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 but 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 for 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 quality of image. 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 there is no explanation provided. Ada is used to searching for relevant images in picture libraries but has never used this type of DSS, 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 continues to scroll 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 colleague to discuss why they think DSS made the 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 that 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 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 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 visualisations to indicate critical information and highly valuable features 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 or risk (calibrate the perception of risk).

The second journalist she turns to, Marc, is pulling 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 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 own 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 but 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, certain celebrities, the system generates a selection of recommended images with metadata attached including the photographers description, date taken and copyright information. As he does this, he finds himself questioning why each image has been picked as it is not always clear the connection between terms he has used and the image or how he can assess the accuracy of each suggestion. To double-check, he searches online for details about the celebrities depicted in the images in order to get a better sense of their relevance and importance and to write captions for those he chooses to include. Leila notes that this is a situation of low journalistic 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 that 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 accompanied explanations/suggestions) may enable practice, learning and potential development of expert skills. Because he is under low time-pressure, it is advisable to enable 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 development of expert skills. Leila decides 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 and allow Marc to further question the output (facilitate 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 back on his work process, he realises that only white celebrities were recommended by the system and wonders in what way and to what extent the system's recommendations impacted his decisions about which images to include.

7.2. Scenario 2: Data-mining and visualisation system

Interface designer Jo has been tasked with considering how to build explanations into a new ML-driven data mining and visualisation tool for investigative journalists that finds and displays connections between data. She is sat with Salim, who is 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, but 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. In high-risk, low time-pressure situations like these, experts are highly motivated and are 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, Salim is under low time-pressure, which means supporting slow analytical thinking would be beneficial, by allowing him to manipulate the data and make comparisons (e.g., interactive models, 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 decision-making process, b) support flexible 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 decides 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 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 an option to explore multiple variants within categories using refinement tools (exploration of 'noise' in data).

These 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. But it also indicates there is a limit to what can be achieved with the tool when situational profiles can be dynamic – how can explanations adapt and respond to the combination of cognitive aspects we have identified?

8. Conclusion

8.1. Future opportunities and limitations

Having established the ERT tool and design approaches to guide expert users, the next steps for developing it as a tool for improved AI intelligibility, in the service of more accountable sociotechnical systems, involves application and research in the wild. We envisage in-situ testing, development and engagement across different contexts, as well as evaluation.

Though it is out of scope in this article to elaborate in detail how a user-interface would look and work, we aim to support and strengthen the applicability of the ERT framework by conducting further research consulting interface designers and ML experts. Through this participatory approach, we hope to distil further concrete design approaches and explainability techniques that could be linked to the segments of the ERT framework. In the future we also hope to explore the effect of using explainability that supports naturalistic decision-making strategies on the development of expertise.

However, it is important to recognise the pressures facing organisations that may wish to develop their own 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 human factors, 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 trying to replicate human decision-making strategies, we should also be careful to not transfer human biases into algorithm-supported decisions. Although expert decision-making has unique qualities, all humans are susceptible to the use of various imperfect heuristics. In our review we touched on several of these heuristics, 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 that we have not covered here.

8.2. Contribution summary

This paper identifies several core ways in which the design of algorithmic decision support systems fail to meet the cognitive needs of domain experts and argues for the importance of psychology-driven differentiated intelligibility support/design for expert users. In support of this it highlights three core dynamics that most heavily influence the manner and type of explanation required, applies these to a worked example of an expert domain, and makes indicative suggestions for how designers might begin to think around the issues.

Our work provides a novel approach to explainability in the decision-making context based on tailoring explanations to facilitate the use of experts' naturalistic decision-making and sensemaking strategies. Drawing from human factors and cognitive psychology we illustrate how these strategies can be determined by the key factors of human expertise, risk and time. We develop a domain-agnostic conceptual framework that could inform explainability researchers and guide interface designers by providing information about key decision-making strategies to be supported in any given decision-making scenario.

The ERT framework was found to be practical and fit for purpose in the journalistic application area and potential in other domains. 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. This will be of relevance to both explainability researchers and practitioners.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge support from EPSRC grants EP/

T517501/1 and EP/S035362/1. The authors would like to thank the anonymous reviewer for the constructive comments that greatly contributed to improving the final version of the manuscript.

References

Webster's new world dictionary. (p. 168). (1968) (p. 168). Cleveland, OH: The World Publishing Company.

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. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 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 : practical innovations, open solutions, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560). https://doi.org/10.1145/3233547.3233667

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., et al. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 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

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

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (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

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

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., et al. (2021). Does the whole exceed its parts? The effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3411764.3445717

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., et al. (2020). *Flexible and context-specific AI explainability: A multidisciplinary approach*. https://doi.org/10.2139/ssrn.3559477. https://doi.org/.

Beckett, C. (2019). New powers, new responsibilities – a global survey of journalism and artificial intelligence. *LSE Polis Report*. https://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities.

Belair-Gagnon, V., & Holton, A. E. (2018). Boundary work, interloper media, and analytics in newsrooms: An analysis of the roles of web analytics companies in news production. *Digital Journalism, 6*(4), 492–508. https://doi.org/10.1080/21670811.2018.1445001

Benner, P., Tanner, C., & Chesla, C. (1992). From beginner to expert: Gaining a differentiated clinical world in critical care nursing. *ANS. Advances in nursing science, 14*(3), 13–28. https://doi.org/10.1097/00012272-199203000-00005

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

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., et al. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648–657). https://doi.org/10.1145/3351095.3375624

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta psychologica, 81*(3), 211–241.

Brennen, A. (2020). What do people really want when they say they want" Explainable AI?" We Asked 60 Stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). https://doi.org/10.1145/3334480.3383047

Breznitz, S. (1989). Information induced stress in humans. *Molecular biology of stress* (pp. 253–264). New York: Liss.

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. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 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

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics* (pp. 160–169). https://doi.org/10.1109/ICHI.2015.26

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

Camerer, C. F., Johnson, E., Rymon, T., & Sen, S. (1993). Cognition and framing in sequential bargaining for gains and losses. *Frontiers of game theory, 104*, 27–47.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721–1730). https://doi.org/10.1145/2783258.2788613

Cesna, M., & Mosier, K. (2005). Using a prediction paradigm to compare levels of expertise and decision making among critical care nurses. In H. Montgomery, R. Lipshitz, & B. Brehmer (Eds.), *How professionals make decisions* (pp. 107–117). CRC Press.

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., et al. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12). https://doi.org/10.1145/3290605.3300789

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency* (pp. 134–148).

Christin, A. (2020). *Metrics at work: Journalism and the contested meaning of algorithms*. Princeton University Press.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev., 89*, 1–34.

Coddington, M. (2015). Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital journalism, 3*(3), 331–348.

Coderre, S., Mandin, H. H. P. 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

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

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies, 2015*(1), 92–112. doi: ArXiv:1408.6491.

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series. Artificial Intelligence for Human-Robot Interaction AAAI Technical Report FS-17-01*.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). https://doi.org/10.1145/3313831.3376638. https://doi.org/.

Diakopoulos, N. (2014). *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*.

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

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Cambridge, Massachusetts: 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.

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

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

Dörner, D. (1990). The logic of failure. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 327*(1241), 463–473. https://doi.org/10.1098/rstb.1990.0089

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. .arXiv preprint arXiv:1702.08608

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: Free Press.

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.

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

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological review, 85*(5), 395. https://doi.org/10.1037/0033-295X.85.5.395

Elwyn, G., Scholl, I., Tietbohl, C., Mann, M., Edwards, A. G., Clay, C., et al. (2013). Many miles to go…": A systematic review of the implementation of patient decision support interventions into routine clinical practice. *BMC medical informatics and decision making, 13*(2), 1–10.

Ericsson, K. A., & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. Cambridge: Cambridge University Press.

Eva, K. W., & Cunnington, J. P. (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

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

Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., et al. (2020). Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 69–78). https://doi.org/10.1145/3397271.3401051

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

Garcia, F. J. C., Robb, D. A., Liu, X., Laskov, A., Patron, P., & Hastie, H. (2018). Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 99–108).

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/DSAA.2018.00018

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 90–99). https://doi.org/10.1145/3287560.3287563

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

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

Gutierrez-Lopez, M., Missaoui, S., Makri, S., Porlezza, C., Cooper, G., & Macfarlane, A. (2019). Journalists as design partners for AI. *In Workshop for accurate, impartial and transparent journalism: Challenges and solutions. CHI 2019*.

Hansen, M., Roca-Sales, M., Keegan, J. M., & King, G. (2017). *Artificial intelligence: Practice and implications for journalism*. Columbia University. https://doi.org/10.7916/D8x92PRD

Hanusch, F. (2017). Web analytics and the functional differentiation of journalism cultures: Individual, organizational and platform-specific influences on newswork. *Information, Communication & Society, 20*(10), 1571–1586. https://doi.org/10.1080/1369118X.2016.1241294

Helberger, N. (2019). On the democratic role of news recommenders. *Digital Journalism, 7*(8), 993–1012.

Helberger, N., Eskens, S.J., .van Drunen, M.Z., .Bastian, M.B., .& Möller, J.E. (.2019). *Implications of AI-driven tools in the media for freedom of expression. Paper presented at Artificial intelligence – Intelligent politics: Challenges and opportunities for media and democracy, Nicosia, Cyprus*. https://hdl.handle.net/11245.1/64d9c9e7-d15c-4481-97d7-85ebb5179b32.

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., Dhurandhar, A., Mojsilović, A., et al. (2019). TED: Teaching AI to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 123–129). https://doi.org/10.1145/3306618.3314273. https: //doi.org/.

Hoffman, R. R. (1998). How can expertise be defined? Implications of research from cognitive psychology. In R. Williams, W. Faulkner, & J. Fleck (Eds.), *Exploring expertise* (pp. 81–100). Palgrave Macmillan.

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.

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining explanation, part 4: A deep dive on deep nets. *IEEE Intelligent Systems, 33*(3), 87–95. https://doi.org/10.1109/MIS.2018.033001421

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science (New York, N.Y.), 288*(5465), 531–532. https://doi.org/10.1126/science.288.5465.531

Jones, B., Jones, R., & Luger, E. (forthcoming). AI 'everywhere and nowhere': Addressing the intelligibility problem in public service journalism. Digital Journalism.

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.

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

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. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3313831.3376219

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological Review, 103*(4), 687. https://doi.org/10.1037/0033-295X.103.4.687

Khairat, S., Marc, D., Crosby, W., & Sanousi, Al (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

Klein, G. A. (2003). Intuition at work: Why developing your gut instincts will make you better at what you do. *Currency/Doubleday*.

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., & Chase, V. M. (1998). Sources of power: How people make decisions. *Nature, 392*(6673), 242. -242.

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

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

Kube, A., Das, S., & Fowler, P. J. (2019). Allocating interventions based on predicted outcomes: A case study on homelessness services. *In Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 622–629. https://doi.org/10.1609/aaai.v33i01.3301622

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 126–137). https://doi.org/10.1145/2678025.2701399

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., et al. (2019). Human evaluation of models built for interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*(1), 59–67.

Langer, E. J., & Abelson, R. P. (1983). *The psychology of control*. SAGE Publications. Incorporated.

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. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3365–3376). https://doi.org/10.1145/3025453.3025884

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). https://doi.org/10.1145/3313831.3376590

Liebhaber, Michael J., & Feher, Bela (2002). Air threat assessment: Research, model, and display guidelines. *SPACE AND NAVAL WARFARE SYSTEMS COMMAND SAN DIEGO CA*.

Liebhaber, M. J., & Feher, B. (2002). Air threat assessment: Research, model, and display guidelines. *Space And Naval Warfare Systems Command San Diego Ca*.

Lipshitz, R., & Cohen, M. S. (2005). Warrants for prescription: Analytically and empirically based approaches to improving decision making. *Human factors, 47*(1), 102–120.

Lipshitz, R., & Strauss, O. (1997). Coping with uncertainty: A naturalistic decision-making analysis. *Organizational Behavior and Human Decision Processes, 69*(2), 149–163.

Long, D., Capan, M., Mascioli, S., Weldon, D., Arnold, R., & Miller, K. (2018). Evaluation of user-interface alert displays for clinical decision support systems for sepsis. *Critical Care Nurse, 38*(4), 46–54.

Lundberg, S., & Lee, S.I. (.2017). A unified approach to interpreting model predictions. arXiv preprint ar*Xiv:1705.07874*.

Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). *A grounded interaction protocol for explainable artificial intelligence*. arXiv preprint ar*Xiv:1903.02409*.

Maiden, N., Zachos, K., Brown, A., Brock, G., Nyre, L., Nygård Tonheim, A., et al. (2018). Making the news: Digital creativity support for journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.1145/3173574.3174049

Marconi, F. (2020). *Newsmakers: Artificial intelligence and the future of journalism*. Columbia University Press.

Edited by Marek, J., Tangenes, B., & Hellesoy, O. (1985). Experience of risk and safety. In O. Hellesoy (Ed.), *Work environment statfjord field. work environment, health and safety on a north sea oil platform* (pp. 142–174). Oslo: Norwegian University. Edited by.

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. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 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

Naiseh, M., Jiang, N., Ma, J., & Ali, R. (2020). Personalising explainable recommendations: Literature and conceptualisation. In *World Conference on Information Systems and Technologies* (pp. 518–533). Cham: Springer.

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

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

Omodei, M. M., Wearing, A. J., McLennan, J., Elliott, G. C., & Clancy, J. M. (2005). More is better? Problems of self-regulation in naturalistic decision making settings. In

Henry Montgomery, Raanan Lipshitz, & Berndt Brehmer (Eds.), *How professionals make decisions* (pp. 29–42). CRC Press.
Orasanu, J. (1997). Stress and naturalistic decision making - Strengthening the weak links. In R. Flin, E. Salas, M. Straub, & L. Martin (Eds.), *Decision-making under stress: Emerging themes and applications* (pp. 43–66). Routledge.
Orasanu, J. (2005). Crew collaboration in space: A naturalistic decision-making perspective. *Aviation, Space, and Environmental Medicine, 76*(6), B154–B163.
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
Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). *The algorithmic automation problem: Prediction, triage, and human effort*. arXiv preprint arXiv:1903.12220.
Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5–14.
Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Cham: Springer.
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.
Reason, J. (1990). *Human error*. Cambridge university press.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
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, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The cambridge handbook of expertise and expert performance* (pp. 403–419). Cambridge University Press.
Rosson, M. B., & Carroll, J. M. (2003). Scenario-based design. In J. A. Jacko, & A. Sears (Eds.), *Human-Computer interaction: Development process* (pp. 1032–1050). L. Erlbaum Associates Inc.
Rundmo, T. (2001). Employee images of risk. *Journal of Risk Research, 4*(4), 393–404.
Rundo, L., Pirrone, R., Vitabile, S., Sala, E., & Gambino, O. (2020). Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of biomedical informatics, 108*, Article 103479. https://doi.org/10.1016/j.jbi.2020.103479
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
Schaffer, J., Giridhar, P., Jones, D., Höllerer, T., Abdelzaher, T., & O'donovan, J. (2015). Getting the message? A study of explanation interfaces for microblog data analysis. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 345–356). https://doi.org/10.1145/2678025.2701406
Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 240–251).
Schmitt, J., & Klein, G. (1999). *A recognitional planning model*. Fairborn OH: Klein Associates Inc.
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. SAGE Publications.
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., & van Merrienboer, J. J. (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. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (pp. 263–268). https://doi.org/10.1145/3393914.3395862
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
Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). *Automation bias and errors: Are crews better than individuals?*. In . *10* pp. 85–97). https://doi.org/10.1207/S15327108IJAP1001_5
Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis: An International Journal, 24*(2), 311–322.
Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., et al. (2020). No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376624
Speier, C., & Morris, M. G. (2003). The influence of query interface design on decision-making performance. *MIS quarterly*, 397–423.
Srinivasan, R. M., & Chander, A. (2020). Generating user-friendly explanations for loan denials using generative adversarial networks. *Fujitsu Technical Review*, 1–6.
Stark, J. A., & Diakopoulos, N. (2016). Towards editorial transparency in computational journalism. In *Computation+ Journalism Symposium* (Vol. 5).
Sterman, J. D., & Sweeney, L. B. (2004). Managing complex dynamic systems: challenge and opportunity for. In Henry Montgomery, Raanan Lipshitz, & Berndt Brehmer (Eds.), *How professionals make decisions*. CRC Press.
Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism, 7*(8), 1076–1097. https://doi.org/10.1080/21670811.2019.1630289
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
Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). Investigating Human+ Machine Complementarity for Recidivism Predictions. arXiv preprint 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 preprint arXiv:*1806.07552*.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.), 185*(4157), 1124–1131.
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
VanBerlo, B., Ross, M. A., Rivard, J., & Booker, R. (2021). Interpretable machine learning approaches to prediction of chronic homelessness. *Engineering Applications of Artificial Intelligence, 102*, Article 104243. https://doi.org/10.1016/j.engappai.2021.104243
Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–14). https://doi.org/10.1145/3173574.3174014
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
Wanner, J., Heinrich, K., Janiesch, C., & Zschech, P. (2020). How much AI do you require? *Decision Factors for Adopting AI Technology*.
Weld, D. S., & Bansal, G. (2018). *Intelligible artificial intelligence*. ArXiv e-prints.
Whalen, J. (1995). Expert systems versus systems for experts: Computer-aided dispatch as a support system in real-world environments. *Cambridge Series on Human Computer Interaction*, 161–183.
Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 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
Wolf, C. T. (2019). Explainability scenarios: Towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 252–257). https://doi.org/10.1145/3301275.3302317
Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.1145/3290605.3300468
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. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12). https://doi.org/10.1145/3290605.3300509
Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data & Society, 6*(2), Article 2053951719868492. https://doi.org/10.1177/2053951719868492
Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307–317). https://doi.org/10.1145/3025171.3025219
Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems, 11*(2), 68–75. https://doi.org/10.1049/iet-its.2016.0208