Int. J. Human–Computer Studies 169 (2023) 102941

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

How the different explanation classes impact trust calibration: The case of clinical decision support systems

Mohammad Naiseh [a,*], Dena Al-Thani [c], Nan Jiang [b], Raian Ali [c]

[a] Faculty of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK
[b] Faculty of Science and Technology, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK
[c] College of Science and Engineering, Hamad Bin Khalifa University, Qatar

ARTICLE INFO

Keywords:
Explainable AI
Clinical decision support systems
Human-AI Interaction
Trust Calibration

ABSTRACT

Machine learning has made rapid advances in safety-critical applications, such as traffic control, finance, and healthcare. With the criticality of decisions they support and the potential consequences of following their recommendations, it also became critical to provide users with explanations to interpret machine learning models in general, and black-box models in particular. However, despite the agreement on explainability as a necessity, there is little evidence on how recent advances in eXplainable Artificial Intelligence literature (XAI) can be applied in collaborative decision-making tasks, i.e., human decision-maker and an AI system working together, to contribute to the process of trust calibration effectively. This research conducts an empirical study to evaluate four XAI classes for their impact on trust calibration. We take clinical decision support systems as a case study and adopt a within-subject design followed by semi-structured interviews. We gave participants clinical scenarios and XAI interfaces as a basis for decision-making and rating tasks. Our study involved 41 medical practitioners who use clinical decision support systems frequently. We found that users perceive the contribution of explanations to trust calibration differently according to the XAI class and to whether XAI interface design fits their job constraints and scope. We revealed additional requirements on how explanations shall be instantiated and designed to help a better trust calibration. Finally, we build on our findings and present guidelines for designing XAI interfaces.

1. Introduction

Recent advances in machine learning have increased the adoption of human-AI collaborative decision-making tools in safety-critical applications such as healthcare systems (Bayati et al., 2014; Caruana et al., 2015) and criminal justice systems (Flores et al., 2016). Combining humans and AI in a collaborative decision-making task is expected to increase the quality of the decision-making (Green and Chen, 2019; Jacobs et al., 2021). However, recent studies showed that humans frequently make trust calibration mistakes by following incorrect recommendations or rejecting correct ones (Jacobs et al., 2021; Bussone et al., 2015; Zhang et al., 2020). In the context of Human-AI environments, trust calibration is defined as an appropriate trust judgement made by humans regarding the current state of AI capabilities and as a successful assessment of whether to follow or reject AI recommendations (Lee and See, 2004; Buçinca et al., 2021). It has been identified as a key design goal for safe and effective Human-AI collaboration (Amershi et al., 2019).

One approach to successful trust calibration is eXplainable AI (XAI) which refers to an AI component that explains AI recommendations to humans receiving them (Naiseh et al., 2021c). Explainability has been identified as a requirement to promote reliability and trust in the AI output and also to ensure humans remain in control (Holzinger, 2021). Explanations can help a human operator understand the AI rationale and reasoning as well as decide when to accept an AI-based recommendation or reject it (Yang et al., 2020; Lai and Tan, 2019; Cai et al., 2019). Explanations are also a critical factor when establishing liability and accountability for the final decisions (Hagras, 2018; Dazeley et al., 2021). Two main streams emerged in the fast-growing XAI research. The first suggests new interpretable machine learning models that are mathematically explainable and transparent, which can compete with black-box models' performance (Arrieta et al., 2020). On the other hand, model-agnostic approaches suggest interpreting any machine learning model, whether interpretable or black-box models, by analysing the

* Corresponding author.
E-mail address: m.naiseh@soton.ac.uk (M. Naiseh).

https://doi.org/10.1016/j.ijhcs.2022.102941
Received 1 May 2022; Received in revised form 30 September 2022; Accepted 1 October 2022
Available online 8 October 2022

model and uncovering its decision rationale (Hohman et al., 2019). Examples of approaches utilising model-agnostic techniques are Interpretable Local Surrogates, Occlusion Analysis, Integrated Gradients and Layerwise Relevance Propagation (Samek et al., 2021). A model-agnostic approach is motivated by preserving the models' confidentiality, increasing the cost-efficiency of generating the explanation, and increasing its usability (Feng and Boyd-Graber, 2019; Zhang et al., 2020; Sokol and Flach, 2019). Different model-agnostic methods may generate explanations with distinct explanation outputs, but they may vary in their performance, fidelity and completeness of the underlying AI model (Arrieta et al., 2020). This paper refers to the family of model agnostic methods that generate distinct explanation output as an XAI class. XAI class can provide answers to similar users' questions (Carvalho et al., 2019; Liao et al., 2020).

While many studies emphasised the requisite of explainability to support trust calibration (Lai and Tan, 2019; Naiseh et al., 2021c; Zhang et al., 2020), several recent studies found little evidence that the XAI class has a significant impact on trust calibration (Jacobs et al., 2021). Many reasons could have contributed to this effect. One is the complex nature of trust. According to the human-computer trust (HCT) model (Madsen and Gregor, 2000), trust is formed in two dimensions: cognition-based trust and affect-based trust. Cognition-based trust is based on humans' intellectual perceptions of AI reasoning, whereas affect-based trust is based on humans' emotional responses to AI systems. In fact, several studies suggest that XAI class output (Wang et al., 2019), its level of transparency (Kulesza et al., 2013; Guesmi et al., 2021) and framing (Narayanan et al., 2018) are all factors that can affect both humans' cognition and affect. Another reason for the effect of XAI class on trust calibration is the nature of humans' cognitive biases (Naiseh et al., 2021c). For instance, under-trust may be resulted from anchoring bias when humans look at only salient features of XAI class and thus judge the quality of the XAI class to be untrustworthy. Similarly, over-trust may result from confirmation bias when humans favour XAI class that is consistent in its output with their beliefs and initial hypothesis.

Although, recent empirical studies have examined the role of XAI class in calibrating users' trust during Human-AI collaborative decision-making tasks (Jacobs et al., 2021; Zhang et al., 2020). The literature is still missing empirical studies including recent advances in XAI literature. In one related research, Dodge et al. (2019) examined the impact of different XAI classes on calibrating the percivied fairness of AI systems. They aimed to determine which XAI class can help participants identify fair AI decisions. They found that Local explanations seem to be more useful than Global explanations when used to explain an unfair model's decisions, thus more effectively calibrating people's fairness judgement. Another study by Zhang et al. (2020) also studied the effect of presenting AI confidence scores and local explanations on users' trust calibration. They found that presenting confidence score to end-users can enhance trust calibration. Another limitation of recent studies that explored the impact of XAI class on trust calibration is that they frequently limited their studies to approaching participants that are unfamiliar with the Human-AI task. Although they boosted participants' knowledge and familiarity of the Human-AI task by introducing a training task, we argue that measuring trust calibration and its relation to XAI class may require expert users with the task to observe the effect during a lived experience and on a fine-grained level. Different from earlier work, we explore the impact of four XAI classes (Local, Example-based, Counterfactual, and Global explanations) on trust calibration during Human-AI collaborative decision-making tasks. We study this effect using a clinical decision support system with experts from the medical domain

In this paper, we endorse the same postulate in Zhang et al. (2020) that calibrating user trust requires different research from the one focusing on increasing users' trust in AI. Inspiring trust can be done without necessarily improving users' mental models and beliefs of the true AI capability and limitations. On the other hand, calibrated trust may need more engagement from end-users to understand the AI system's reasoning (Naiseh et al., 2021c). We study the effect of four distinct XAI classes (Local, Example-based, Counterfactual, and Global explanations) on trust calibration during a collaborative Human-AI task. We hypothesise that different XAI classes can affect trust calibration differently when using an AI-based decision support system. Our research method is based on a within-subject study followed by semi-structured interviews. We developed a classification tool that helps medical practitioners in screening chemotherapy prescriptions, i.e., an AI provides recommendations to either accept the prescription or reject it, representing a high-stakes application domain. Taking clinical decision support systems as an exemplar and focusing on the four state-of-the-art model-agnostic XAI classes, our contribution has two parts:

- A quantitative evaluation of how explanations belonging to the different XAI classes affect users' trust calibration during Human-AI collaborative decision-making tasks.
- Recommendations for XAI interface design to help improve trust calibration using data collected from interviewing medical practitioners.

The findings of our study are intended to provide a richer understanding of the main needs of users from explanations of their different classes. We also aim at broadening discussions on explainable AI for collaborative decision-making and paving the way for more research on how to customise and contextualise explanations so that they fit users' needs and expectations of each of their different classes. Compared to our previous work (Naiseh et al., 2021a; Naiseh et al., 2021b), this study adds a quantitative evaluation of different XAI classes and their effect on users' trust calibration. It also elicits XAI interface requirements correlated with particular XAI classes. In our previous work, we took a broad view of XAI without focusing on the differences between XAI classes. We first explored what errors users make while interacting with XAI interfaces (Naiseh et al., 2021b). We then conducted an explanatory study to discover potential design solutions to mitigate users' errors (Naiseh et al., 2021a).

The paper is structured as follows. Section 2 describes the research method, including the sample, material, and instruments used in our study. In Section 3, we present our analysis results and findings. Finally, in Section 4, we discuss the implications of the findings on future research and development in the field and conclude the paper.

2. Research method

Our study investigates the effect of XAI class on trust calibration during a Human-AI collaborative decision-making task. Through both quantitative and qualitative research, we aim to answer the following questions:

- RQ1. How do different XAI classes affect users' trust calibration?
  - RQ1.1 Do XAI classes affect users' judgments of trust (increase or decrease)?
  - RQ1.2 Are XAI classes seen differently in enabling trust calibration? i.e., Are different XAI classes seen differently in affecting the performance of the Human-AI team?
- RQ2. What are the users' requirements for interfaces used for delivering explanations of different XAI classes?

2.1. Human-AI task description

Chemotherapy screening prescription is a process that practitioners in a clinic follow to ensure that a prescription is prescribed for its clinical purpose and fits the patient's profile and history. We chose the use case of screening prescription as it reflects an everyday decision-making task performed collaboratively between humans and the AI. The main

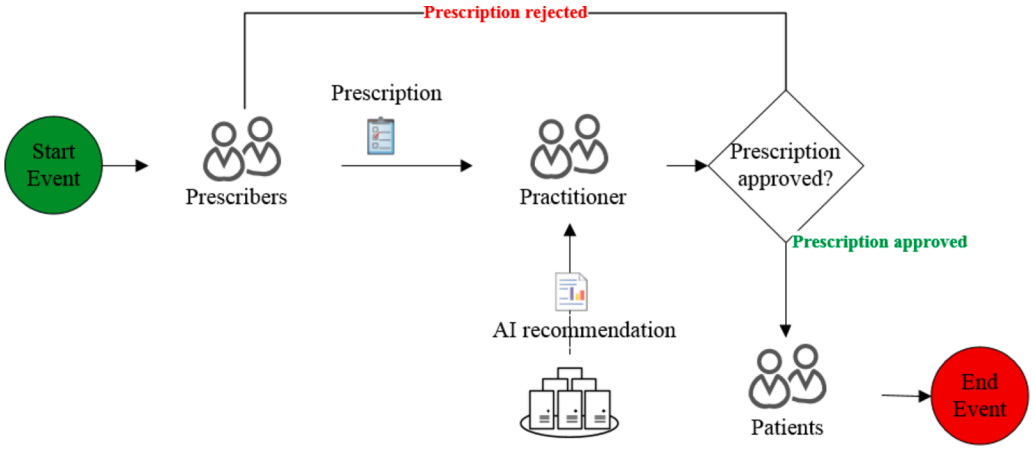

Fig. 1. Workflow for prescription screening aided by AI-based decision-making tool.

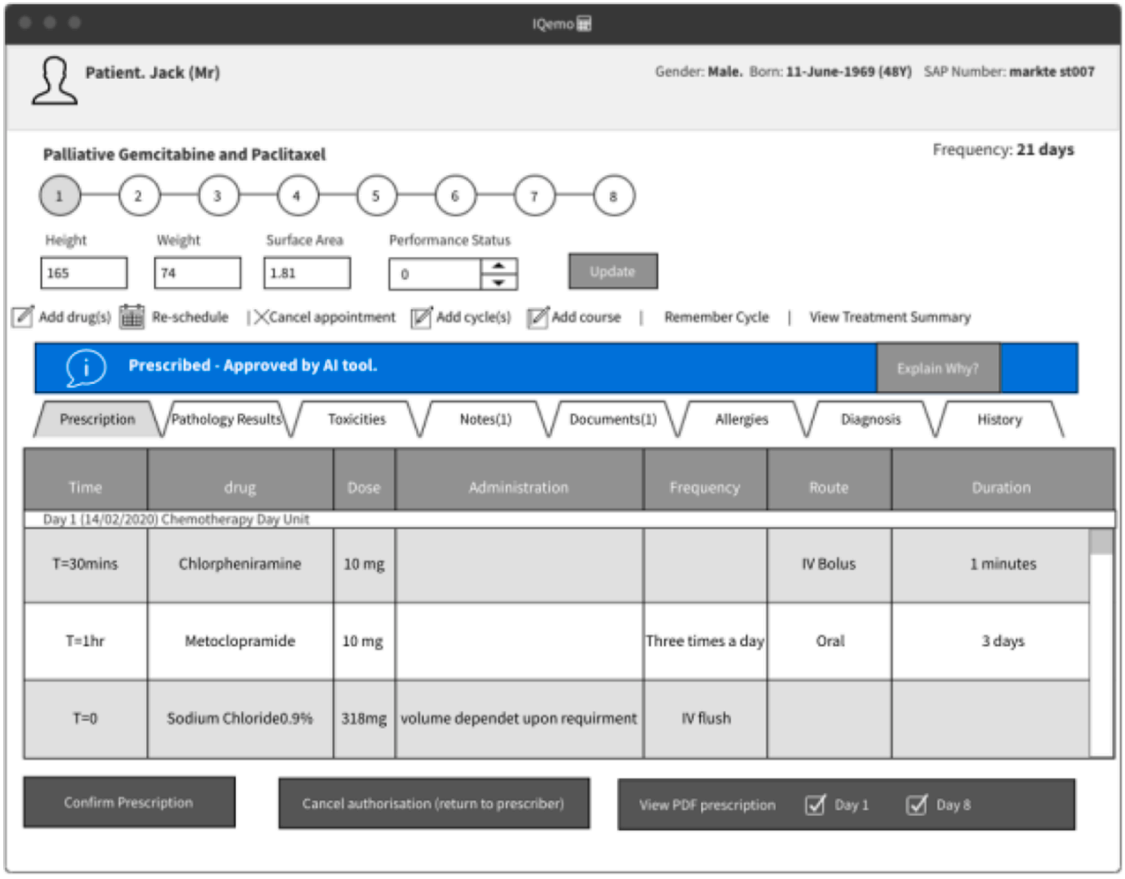

Fig. 2. A sample of prescribing system interface supported with AI recommendations.

workflow of the AI-based prescribing system is shown in Fig. 1.

To help our investigation, we designed an AI-supported decision-making mock-up that classifies prescriptions into confirmed or rejected. We designed the mock-up based on templates and interfaces familiar to our participants in their everyday decision-making tasks (See Fig. 2).

Our mock-ups mimic a web-based tool and are meant to simulate user experience when working on an actual system. As the practitioner clicks on a prescription, the tool shows the patient profile, the recommendation from the AI-supported decision-making tool (accepted or rejected), and an explanation. The explanation can help the practitioner understand the AI rationale of why the prescription should be accepted or rejected. The full material used in our study can be found in Appendix A, B and C.

2.2. Explanation classes

The taxonomy of XAI classes has been adopted from a recent XAI survey (Liao et al., 2020). We choose this classification as it classifies XAI algorithms based on their output in which end-users can recognise the difference between them. A sample of the interfaces representing each of XAI classes, that we showed to our experiment participants, can be found in Appendix B. These XAI classes were used in our mock-up tool.

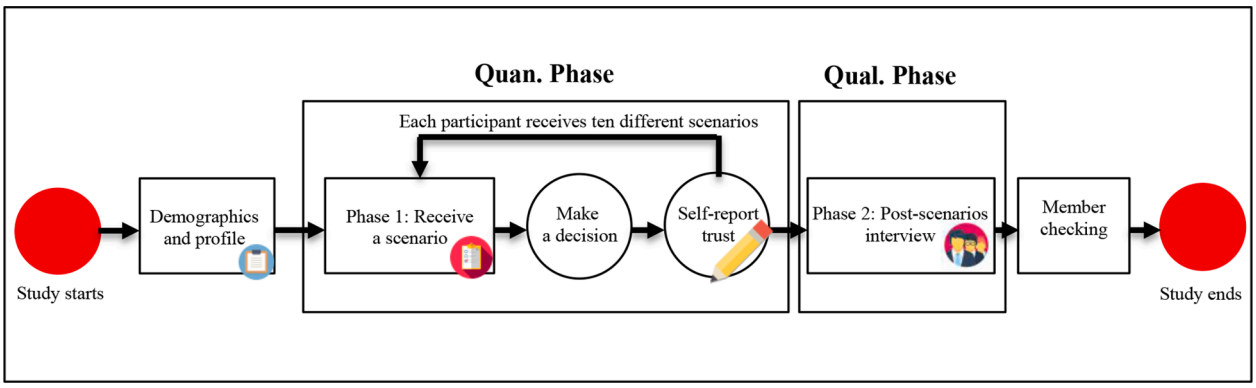

Fig. 3. Study workflow.

1 *Local explanations*: The explanation justifies the AI reasoning at the recommendation level; this can be done either by quantifying the contribution value for each input data feature to the recommendation (Ribeiro et al., 2016) or generating local rules or decision trees of a recommendation (Guidotti et al., 2018). We restricted our study to one type of local explanation which quantifies the importance of each input data feature.
2 *Example-based*: Given the AI-based recommendation, the AI justifies its decision by providing examples from that dataset with similar characteristics (Cai et al., 2019). For example, AI suggests rejecting this prescription because patient A history was similar to patient B.
3 *Counterfactual*: Given the AI-based recommendation, the AI answers the users' questions "*what-if*" to observe the effect of a modified data feature on the recommendation (Sokol and Flach, 2019). For instance, the AI suggests rejecting the prescription because Platelet Count= 60; however, if the Platelet Count were >= 75, the prescription would have been confirmed.
4 *Global explanations*: The explanation of an AI-based recommendation attempts at explaining the overall logic of the black-box model (Henelius et al., 2014; Wu et al., 2020). This includes presenting the weights of different data features as decision trees, rules, or ranking styles. Our study setting included only ranking styles of presenting the weights of data features.

2.3. Study design

The study was a within-subject design followed by follow-up interviews. We manipulated the XAI class (No explanation, Local, Example-based, Counterfactual, Global explanations) and the recommendation outcome (correct and incorrect recommendations). As a result, we had ten different conditions. Each XAI class appeared in two conditions: correct AI-based recommendations and incorrect ones. We developed ten patient scenarios and interfaces to cover the study conditions. Examples of AI recommendations and explanations of different classes can be found in Appendix B. XAI classes were randomly assigned to patient scenarios to eliminate the carryover effect (Louthrenoo et al., 2007) and the potential effect of accidental bias, i.e., having an XAI class with a specific patient scenario. When designing the scenarios, we choose to manipulate the recommendation outcome to simulate a diversity of conditions that the practitioners could face in real-world scenarios where trust calibration errors could happen, e.g., imperfect AI due to the low sample size or recent data. Each participant completed ten different Human-AI tasks. For each participant, we used a random sequence of ten human-AI tasks, i.e. a scenario followed up by questions. The scenarios were diverse and the participants were presented with correct and incorrect recommendations. This has helped eliminate the learning effect on the participants. We chose the first Human-AI task that included a correct recommendation to help the participants gain confidence and engage more with the study (Lee and See, 2004; Marshall, 2003).

The patient scenarios presented to our participants were hypothetical scenarios designed in collaboration with a medical oncologist, i.e., there was no actual AI model. We designed the scenarios to be clear and challenging, and not trivial so that recommendations, explanations, and trust calibration were all substantial processes. This ultimately helped to put our participants in a realistic setting: exposing them to an imperfect AI-based recommendation and its explanations where trust calibration is critical to the task and where errors in that process are possible. We validated the material with a medical oncologist focusing on the border cases that need an investigation from the participants in the actual study. We tested the material and activities with two participants and refined them to optimise their fulfilment of these criteria. Although our scenarios and explanations were not generated via an AI-based model, participants were informed that they are receiving recommendations from an AI-based model and asked to either follow or reject AI recommendations.

2.4. Study procedure and data collection

To answer our research questions, we conducted a study of two phases (Quantitative and Qualitative). We provide an overview of the study phases in Fig. 3.

2.4.1. Quantitative phase

This phase was meant to answer RQ1. The phase involved 41 medical practitioners who have experience in screening prescriptions (Doctors and pharmacists). First, the participants were briefed about the study through a participant information sheet. They were then asked to sign a contest form. Participants were also asked several questions about themselves, such as their experience in chemotherapy prescribing (Appendix A). Personal attributes, including skills and years of experience diversity, should help in covering different types of issues. For example, novice medical experts might raise more questions than those who have more experience in the task since they faced similar cases in the past and can cross-check with other sources. Although we recognise that personal differences, e.g. in agreeableness and conscientiousness (Barrick and Mount, 1991), can play a role, this was beyond the scope of our study. Then, each participant had to complete ten different screening prescription Human-AI tasks. Participants were asked to make decisions considering the patient profile, the recommendation and the explanations and whether to follow the AI-based recommendation if they see it as correct or reject it if they see it as incorrect. Each participant spent 15–20 min completing this stage. The study workflow is described in Fig. 3. After completing a Human-AI task, participants were asked to complete cognitive-based trust scale (Madsen and Gregor, 2000). Each of the participants completed ten Human-AI tasks, which resulted in 410 completed tasks. The following sections describe our measures to answer RQ1.

Table 1
Population details.

Variable	Value	$N = 41$	%
Age	20–30	22	54%
	30–40	11	28%
	40–50	8	18%
Gender	Male	26	65%
	Female	15	35%
Role	Doctors	26	62%
	Pharmacists	15	38%
Prescription Screening Experience	<5	15	35%
	5–10	12	30%
	10–15	9	21%
	>15	5	14%

2.4.2.1. Trust calibration measurements. Trust calibration in automation is defined as the individuals' adjustment of the level of trust they put into the objective automation capabilities and performance with the aim of avoiding under-trusting and over-trusting it (Lee and See, 2004). Considering the challenge of measuring trust calibration, which is a complex psychological construct (Lee and See, 2004), our study focuses on trust calibration by observing the outcome of the interaction between the Human and AI. For the scope of this study, we measure the contribution of an XAI class on trust calibration in two ways: subjective and objective. First, subjective measurements were used by following the cognitive-based trust scale proposed in (Madsen and Gregor, 2000) which quantifies the extent to which the XAI interface helps users to understand, rely on and perceive the technical competency of the AI. Specifically, self-reporting cognitive-based trust measures were used in this study to observe whether an XAI class helped in increasing or decreasing trust during the sessions (RQ1.1). We then look at behavioural indicators of trust calibration as an objective measurement (Wang et al. 2016) based on whether the participants made the right decision. This seems as an indicator of whether an XAI class helped in trust calibration(RQ1.2). This approach to measuring trust calibration was proposed in earlier studies, e.g., (Zhang et al., 2020; Wang et al., 2016; Bussone et al., 2015; Dikmen and Burns, 2022). We further elaborate on our measurements and the rationale behind using them in the following sections.

Self-reporting measures. To understand the contribution of XAI class in affecting (increasing or decreasing) trust during the study, we followed Madsen and Gregor (2000) conceptualisation of human-computer trust which included two main components: cognition-based components and affect-based components. The main difference in their effect is that cognition-based trust is based on human cognitive reasoning to trust another entity and is crucial for maintaining trust calibration, whereas affect-based trust is developed as the relation continues (Nah and Davis, 2002). Furthermore, previous research showed that in critical decision-making scenarios, it is highly likely that cognition-based trust components significantly impact trust calibration in comparison to affect-based ones (McAllister, 1995; Ng and Chua, 2006). Therefore, cognition-based trust was useful to use in our study settings to help us understand whether an XAI class contributed to the increase or decrease of cognitive reasoning to trust when participants interact with the AI. Further, measuring affect-based components was not relevant during our study. It also needs more longitudinal and observational studies, since emotions with an AI need a longer time to be developed. As mentioned earlier, we used the Human-Computer Trust (HCT) scale proposed by Madsen and Gregor (2000) to measure the impact of XAI class on each of the cognition-based trust components. HCT scale has been relatively stable, tested and used in several relevant studies (Yang et al., 2020; Schraagen et al., 2020; Larasati et al., 2020). The scale measures perceived cognition trust based on three main components (*perceived understandability perceived reliability, and perceived technical competence)*. The scale has 15 items (5 items for each cognition-based trust component).

Behavioural indicators. We also collected trust calibration behavioural indicators for each XAI class condition so that we also answer RQ1.2. We utilised three trust calibration behavioural indicators introduced in similar studies (Zhang et al., 2020; Wang et al., 2016; Bussone et al., 2015):

1 *Agreement.* This is a binary variable that indicates whether participants agreed with AI recommendations.
2 *Switch.* This is a binary variable that indicates whether a participant decided to switch from the AI recommendation.
3 *Human-AI performance.* This is a binary variable that indicates whether the collaborative Human-AI task led to a successful decision. i.e., participants agreed with correct recommendations or disagreed with correct ones.

2.4.2. Qualitative phase

The second phase of our study was to answer RQ2. We conducted semi-structured interviews following the guidelines stated in (Oates, 2005). We used interviews as a data collection method to delve into the details and understand the reasoning process and issues faced during the Human-AI collaborative decision-making process (Ericsson and Simon, 1984). We asked our practitioners to elaborate on their experience during their interaction with our mock-up tool to gain as many insights as possible. The interviewer discussed the benefits and the drawbacks of each XAI class in calibrating their trust and experience during the decision-making process. Also, the interviewer asked the participants to express their concerns and difficulties to make an informed decision, given explanations belonging to four XAI classes. Our interview questions were based on three dimensions of cognitive-based trust (Understandability, reliability and technical competence). The full list of questions can be found in Appendix C. After the end of each quantitative phase, participants were asked if they would like to take part in a short interview to discuss their experience. We interviewed 16 participants and analysed their interviews. We did not approach and interview more participants since themes and codes resulting from the analysis became eventually repetitive. This practice is aligned with the principles of reaching the saturation point in qualitative data collection (Faulkner and Trotter, 2017). That means that interviewing more participants will unlikely generate new results. This was a reasonable assurance that further qualitative data collection would introduce similar results and would confirm the existing themes.

2.5. Participants

Ethical approval was obtained through Bournemouth University Ethics Committee. Our sample consisted of 41 medical practitioners coming from three different organisations in the UK. All participants were medical experts recruited through email based on their experience in the case study. We recruited participants who had previous experience in the prescription screening task and used clinical decision support systems before. We choose these inclusion criteria to ensure that trust calibration is observed and measured under realistic settings. Further, we have not designed our study to explore whether differences in the participants' profiles and psychometrics matter concerning the effect of XAI classes on trust calibration. To do this we will need to recruit participants in a planned manner so that we ensure we have representatives from each stratum. In our study, we strived to ensure the fitness of the participants for the task they were asked to go through and their similarity in that sense. The purpose of the study did not primarily relate to the level of experience in the domain (medical/pharmaceutical in our case) and we ensure all had enough so we do not end up with a rejection of explanations due to unfamiliarity with the domain of the explanation. All our participants had a similar level of experience in the study task itself, i.e. interacting with XAI interfaces, and none of them was more or less familiar with XAI in general. Details about the population are provided in Table 1.

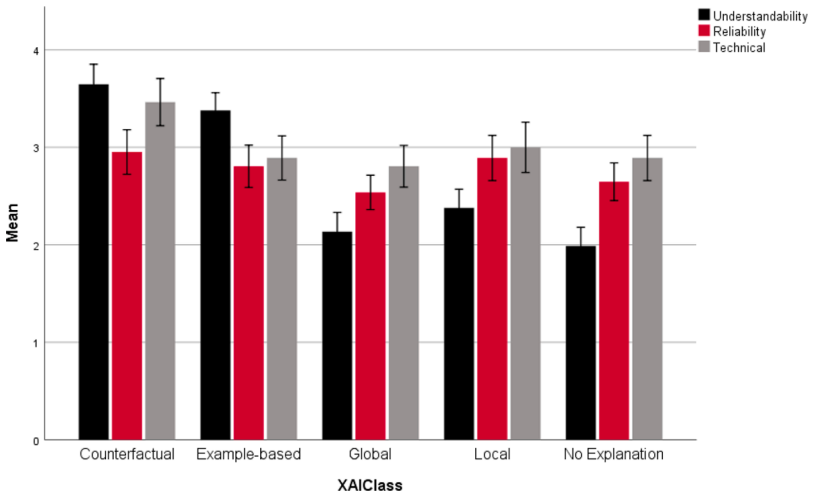

Fig. 4. Mean cognition-based trust components rating per explanation class. Explanation cognition-based trust ratings range from Strongly Disagree (a rating of 1) to Strongly Agree (a rating of 5).

2.6. Data analysis

Two sets of data were collected and used to answer our research questions in this study. The first reflected the trust measurements (self-reporting and behavioural indicators). The second consisted of transcripts of the audio files of the follow-up interviews. We used several statistical tests such as repeated One-way ANOVA to answer our first research question. For qualitative data, we followed the content analysis method with the support of the NVivo[1] tool. To increase the trustworthiness of our qualitative analysis, we applied a member-checking approach and reviewed the analysis of their interviews with them. Member-checking aims to review the analysis report by study participants to ensure that data, interpretations and themes are valid and applicable (Birt et al., 2016). Member checking technique was applied with three participants to validate the analysis and clarify the analysed data where clarification was needed. We used the member-checking method to increase the credibility of our qualitative data and minimise its subjectiveness. Each step of the analysis included several meetings amongst the authors to ensure the correct interpretation of each category and the evidence that supports them. These meetings led to splitting, modifying, discarding or adding categories to ensure that all responses and their contexts were well represented.

3. Findings: explanation classes and trust calibration

In this section, we report on our analysis results answering RQ1 which concerns whether XAI class differs in influencing user trust and trust calibration during a Human-AI collaborative decision-making task. In total, 41 medical experts completed 410 Human-AI tasks.

3.1. Explanation classes impact on user trust (RQ1.1)

Each participant rated their trust (*perceived understandability, perceived reliability, and perceived technical competence*) during the interaction with different XAI classes. The descriptive statistics of participants' ratings are shown in Fig. 4. As a general observation, three trust components were seen differently by our participants, i.e., they were not mutually dependant. For instance, No explanation scenarios were not perceived to be understandable but it has a higher rating in terms of perceived reliability and perceived technical competence. This means that trust indeed is not only about one dimension, and XAI interface design may need to consider each of its components when supporting appropriate trust of the AI. In the following section, we present our results based on each dimension of trust.

3.1.1. Perceived understandability

We applied a repeated measure ANOVA to compare the mean of participants' understandability of the AI between different XAI class conditions (No explanation, Global, Local, Counterfactual and Example-based). Our results show that participants rated how they perceive understandability significantly different between the XAI class conditions [$F(4324) = 63.483$, $p < 0.001$]. Post hoc comparisons using the Tukey honestly significant difference (HSD) test indicated that the mean perceived understandability score in Example-based scenarios ($M = 3.383$, $SD = 0.830$) and Counterfactual scenarios ($M = 3.646$, $SD = 0.935$) was significantly higher than in No explanation ($M = 1.988$, $SD = 0.868$) Local ($M = 2.357$, $SD = 0.873$) and Global ($M = 2.150$, $SD = 0.901$) scenarios.

3.1.2. Perceived reliability

As shown in Fig. 4, clearly, participants rating of how they perceive reliability was steady across multiple XAI classes conditions [No explanation ($M = 2.646$, $SD = 0.880$), Global ($M = 2.537$, $SD = 0.804$), Local ($M = 2.890$, $SD = 1.054$) Example-based scenarios ($M = 2.805$, $SD = 0.987$) and Counterfactual scenarios ($M = 2.951$, $SD = 1.041$)]. Repeated measures ANOVA confirmed this observation and showed no significant difference in participants' perceived reliability of the AI between different XAI class scenarios. We further elaborate on the potential reasons for these indifferences in Section 4, where we report on the issues and needs expressed by the participants considering how they perceived the reliability of the AI. We also discuss the implications of this finding in Section 5.

3.1.3. Perceived technical competence

Results show that the XAI has a significant effect on the users' perceived technical competence of the AI [$F(4324) = 4.815$, $p < 0.001$]. Post hoc comparisons using the Tukey honestly significant difference (HSD) test indicated that the mean perceived technical competency score in Counterfactual scenarios ($M = 3.463$, $SD = 0.1.102$) was significantly higher than in No explanation ($M = 2.890$, $SD = 1.054$) Local ($M = 3.00$, $SD = 1.176$), Global ($M = 2.805$, $SD = 0.974$) and Example-based scenarios ($M = 2.890$, $SD = 1.030$). Our analysis of interview data provided an explanation of that findings when participants mentioned that Counterfactual explanations provided meaningful

[1] https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

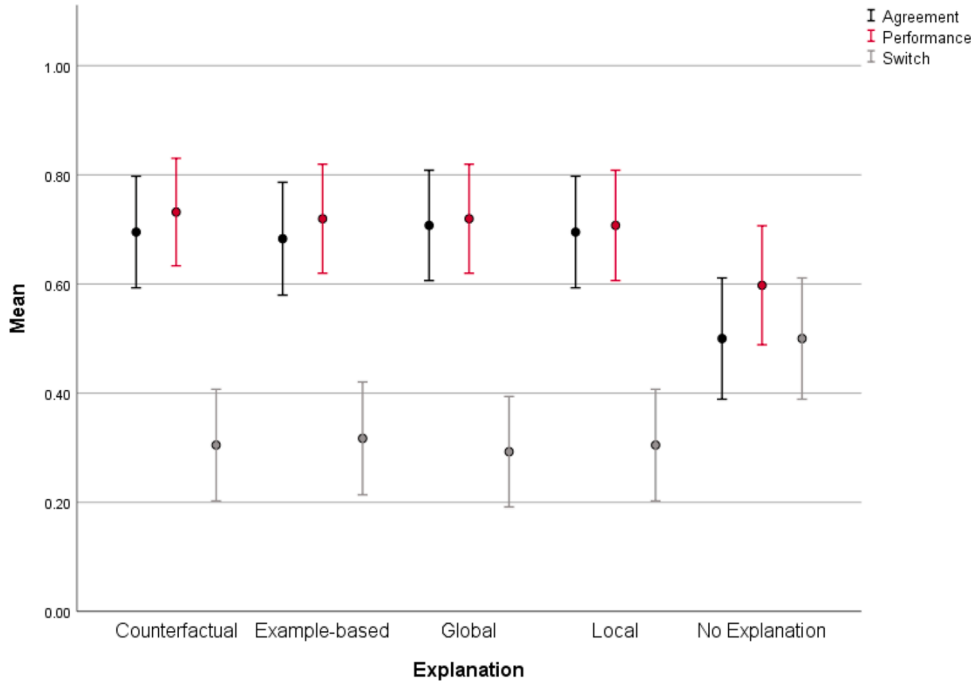

Fig. 5. Three trust collaboration behavioural indicators mean percentage (X-axis represents five XAI classes conditions).

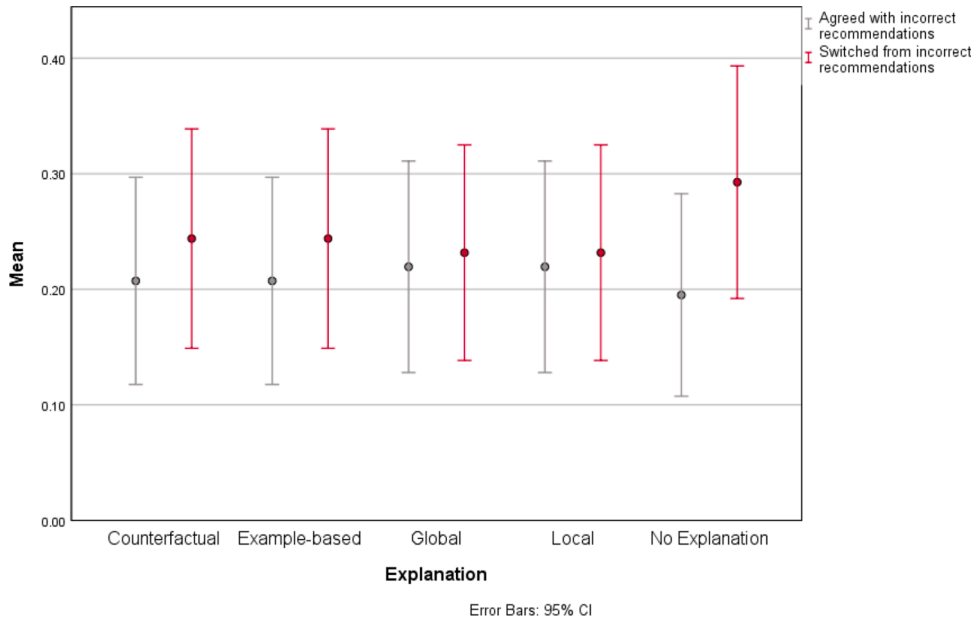

Fig. 6. Switch percentage and agreement percentage for incorrect recommendations across different XAI classes.

knowledge for them to conclude "*Well, I can see how the AI is reasoning about this case, I would say yes the AI is reasonable*" [P12]. Although participants perceived Example-based explanations as an understandable explanation, they did not rate Example-based explanations as significantly competent. P5 commented in that context, "*Examples are beneficial and similar of what we do in the clinic, but it is not a proper explanation … I mean it could be supportive to other explanations … I would expect more casual or correlation relationship between the patient and the AI decision*". In summary, participants considered explanations as competent when the explanation was understandable and showed the rationale behind the specific recommendation and those providing casual patterns at the recommendation level.

3.2. Explanation classes impact on trust calibration process (RQ1.2)

To answer RQ1.2, we looked at three behavioural indicators to examine the effect of XAI class on trust calibration, i.e., whether different XAI classes affected participants' collaborative decision-making.

XAI classes improved the overall Human-AI performance. We plot the comparison of participants' overall performance, average agreement and average switch between XAI class conditions in Fig. 5. Visually, it is clear that when participants were not provided with an explanation, they were more likely to make mistakes and explanations slightly improved the overall collaborative decision-making task. Friedman test confirmed this observation and shows that XAI class significantly affects

Table 2
Issues and needs applicability to explanation classes; (x) applies (-) does not apply.

Issues and needs	Local	Example	Counterfactual	Global
Task-centred explanation	x	x	x	x
Usability	x	x	x	x
Assurances	–	–	–	–
Guidance	x	–	–	x
Tailoring	x	x	x	x
Multi-step explainability	x	x	x	x

Human-AI performance [$\chi2(4) = 19.524$, $p = 0.001$]. Post hoc comparisons using the Wilcoxon signed-rank test indicated that Human-AI performance in Counterfactual ($M = 0.732$, SD = 0.446) Example-based ($M = 0.720$, SD = 0.452), Local explanations. ($M = 0.707$, SD = 0.458) and Global ($M = 0.720$, SD = 0.452) were significantly different than No explanation ($M = 0.598$, SD = 0.793) scenarios. Our results appeared to be consistent with Lai and Tan's (2019) conclusions, which showed that pairing AI recommendations with explanations can improve the Human-AI team.

XAI classes increased participants' agreement with AI recommendations. It is clear from Fig. 5 that participants agreed more with the AI recommendations when explanations of their different classes were provided. Friedman test confirmed this observation and show that XAI class has a significant effect on Human-AI performance [$\chi2(4) = 57.444$, $p = 0.001$]. Post hoc comparisons using the Wilcoxon signed-rank test indicated that Agreement percentage in Counterfactual ($M = 0.695$, SD = 0.463) Example-based ($M = 0.683$, SD = 0.468) Local ($M = 0.695$, SD = 0.463) and Global ($M = 0.707$, SD = 0.458) were significantly different than No explanation ($M = 0.50$, SD = 0.503) scenarios. We also note that agreement and switch measures are complementary and their sum is 100%. That means a significant increase in an agreement between *No explanation* condition and *the four explanations* and also means a significant decrease in *switch percentenge* between *No explanation* and other *the four explanations*. Our findings seem to contradict Zhang et al. (2020) findings when they found that there is no significant effect between users' agreement in three conditions: No explanation, Uncertainty score and Local explanation. However, a closer look at their findings shows that presenting uncertainty scores to participants in their settings could interpret such differences. In other words, the explanation alone with its different classes could have increased our participants' reliance on AI, and the uncertainty score could moderate the over-reliance effect. Our results are consistent with previous research (Bussone et al., 2015), which showed that presenting explanations to end-users could increase participants' over-reliance and facilitate confirmation bias.

XAI classes did not help participants recognise incorrect recommendations. Although XAI classes significantly improved the overall Human-AI team performance, the data suggests that XAI classes did not help participants recognise the incorrect recommendations compared to No explanation scenarios. The average of mistakes during the incorrect recommendation scenarios, i.e., agree with incorrect recommendations, made by participants during XAI classes conditions and No explanation condition was not significantly different [$\chi2(4) = 5.640$, $p = 0.231$]. Fig. 6 compares participants' responses during incorrect recommendations scenarios (Agreeing and switching from incorrect recommendations). It is clear that when the AI recommendation was incorrect participants struggled to decide whether to follow or reject the AI recommendation across all XAI conditions.

We summarise the results from the quantitative analysis in three main points.

- Example-based and Counterfactual explanations had a higher users' perceived understandability than Global, Local and No explanations scenarios and helped our participants to reason about the AI recommendation.
- Users perceived technical competence seemed to be affected by explanation understandability and provide casual patterns at the recommendation level.
- Explanations of its different classes increased the overall performance of the Human-AI team, however, we did not find a significant difference in Human-AI performance when facing incorrect recommendations.

4. Findings: user requirements from xai interfaces

In this section, we discuss our analysis results answering RQ2 concerning the user requirements for better utilisation of interfaces belonging to each XAI class during the Human-AI task. Upon completing the cognitive-based trust scale, participants were interviewed to discuss the main issues they faced during their interaction with the XAI interface and the reasons behind their ratings. Table 2 summarises our interview analysis results.

4.1. Guidance

Participants commented that they needed guidance to interpret Local and Global XAI classes. For instance, P9 mentioned, "*I think it is unfair for AI to explain this way because it just does whatever it was designed to explain for, so it does not give us to see the big picture … I would like to know what it means to have a patient age with 35% influence on the AI decision? and how this could be interpreted for this patient?*". One interpretation of such comments is that Local and Global explanations may require previous technical knowledge to interpret them (Kaur et al., 2020). Our results also revealed that participants misinterpreted these explanations when the interviewer reviewed the explanations with them. For example, P8 commented on the Global explanation encountered during the study, "*I saw that blood test is the influential factor, and I was wondering we should screen prescriptions on that factor only?*". Also, P12 commented on the Local explanation presented during the study: "*I feel this could be biased in some way, so that means the majority of the decisions will be made based on the tumour size*". Participants' interpretations of the potential bias and selectiveness in the explanation could be justified because of participants' unfamiliarity with AI explanations even though they were given a presentation and examples before the study. Furthermore, our results showed that participants who gave low understandability and reliability ratings to Local and Global XAI classes mentioned that the design might need more attention and guidance to help them in interpreting the explanations, e.g., P3 responded, "*I think a tutorial on how to use these explanations would be beneficial*". The previous comments signify that unfamiliarity with some XAI classes can be a significant issue that may decrease users' trust in the AI and potentially lead to trust calibration mistakes. Our results are consistent with previous research that showed XAI methods are mainly designed and used by data scientists and provide little value to other end-users (Kaur et al., 2020; Ras et al., 2018).

4.2. Usability

Although participants identified the usefulness of the explanations to understand the AI rationale, some mentioned that they would not use these explanations in everyday scenarios. One reason for that was related to their concerns of fitting explanations in their workflow, mainly due to the need to process too much information, e.g., P6 stated, "*…. sometimes we are so busy; I won't have that time to validate the AI through its explanation; in my opinion, a simple explanation targeting main patient issues would be enough with an option to investigate more when needed*". In addition, participants felt that explanations could be a burden in their workflow and might cost extra cognitive efforts, e.g., P12 mentioned, "*Does this mean that I have to look at all these factors each time I make a decision?*". Participants suggested that explanations could be better fitted when presented in textual formats as narratives or templates to reduce such a cognitive load and task impediment. For example, P13

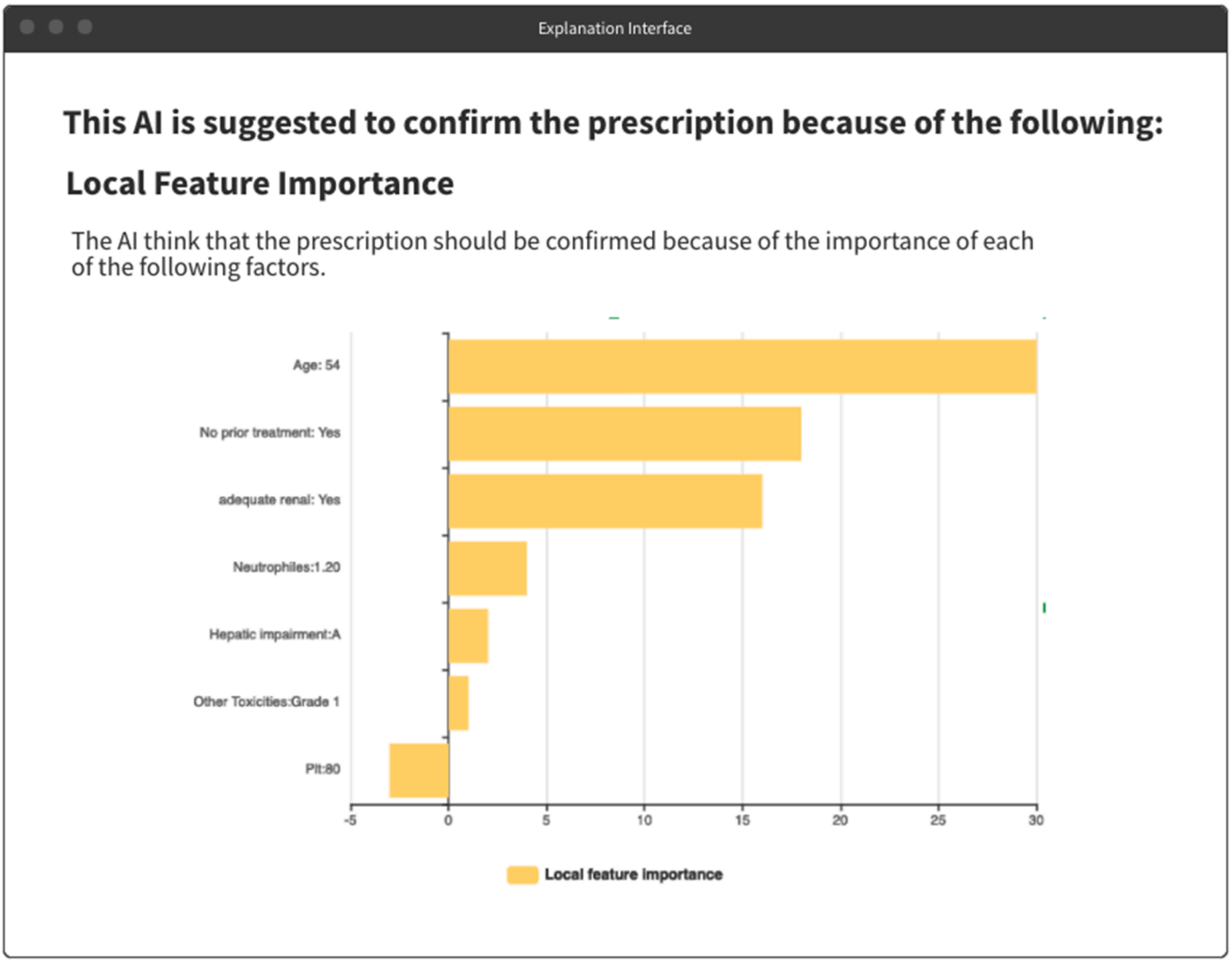

Fig. 7. Local Feature explanation example.

wanted to have a generated narrative that summarises multiple XAI classes. In HCI, this happens when the tool's design leads to a mismatch between the user mental model and the conceptual mental model of the system (Carroll and Olson, 1988). Participants also mentioned that explanations sometimes contained redundant information, and they recommended different ways to make their trust calibration require fewer efforts. For instance, P9 asked to customise the number of data features in Local explanations, "*The average pharmacist does not need to see all these factors that the AI is considering, some of them are just simple rules*". Our observations are aligned with recent studies that showed that long and redundant explanations made participants skip them (Naiseh et al., 2021c) and decreased participants' satisfaction with the explanation (Narayanan et al., 2018). In summary, explanations easiness of use and their modalities could be critical to engaging users with them when participants' time constraints and the difficulty of the task are primary issues. Also, limited time and an unaffordable need for cognition (Petty and Cacioppo, 1986) can be both seen as a contextual disabilities for people who work on pressuring domains. Future work needs to consider the trade-off between effectiveness and usability of the explanation to optimise the Human-AI team performance, e.g., adaptive and personalised user interfaces could be a potential solution direction (Naiseh et al., 2020).

4.3. Task-centred explanation needs

Participants described that all XAI classes did not consider the task needs and constraints, and they were expecting an explanation of the context prior to being presented with the explanation itself. Participants who identified these needs provided low technical competency and reliability rates. One example of task constraint needs was encountered in the Counterfactual explanation scenarios, where the explanation only provided information about changes that could be made to an AI recommendation to change the decision. In our case study, Counterfactual explanation scenario explained the recommendation by modifying the value of the patient blood test but also failed to specify that this modification could trigger risk or require another change in another data feature. Participants raised a concern regarding potential risks behind the explanation as it did not meet their task constraints. For example, in Counterfactual explanation scenarios, explaining the recommendation through a change in a data feature value without identifying the correlation with other data features was perceived as a risky explanation in their decision-making process. This means that making hypothetical scenarios without considering the domain and task constraints, such as the medical domain and patient cases, can result in unrealistic cases; cases that can be incoherent in the values of their variables and lead to additional explanation needs beyond the explaining recommended decision. The main reason for this drawback could be

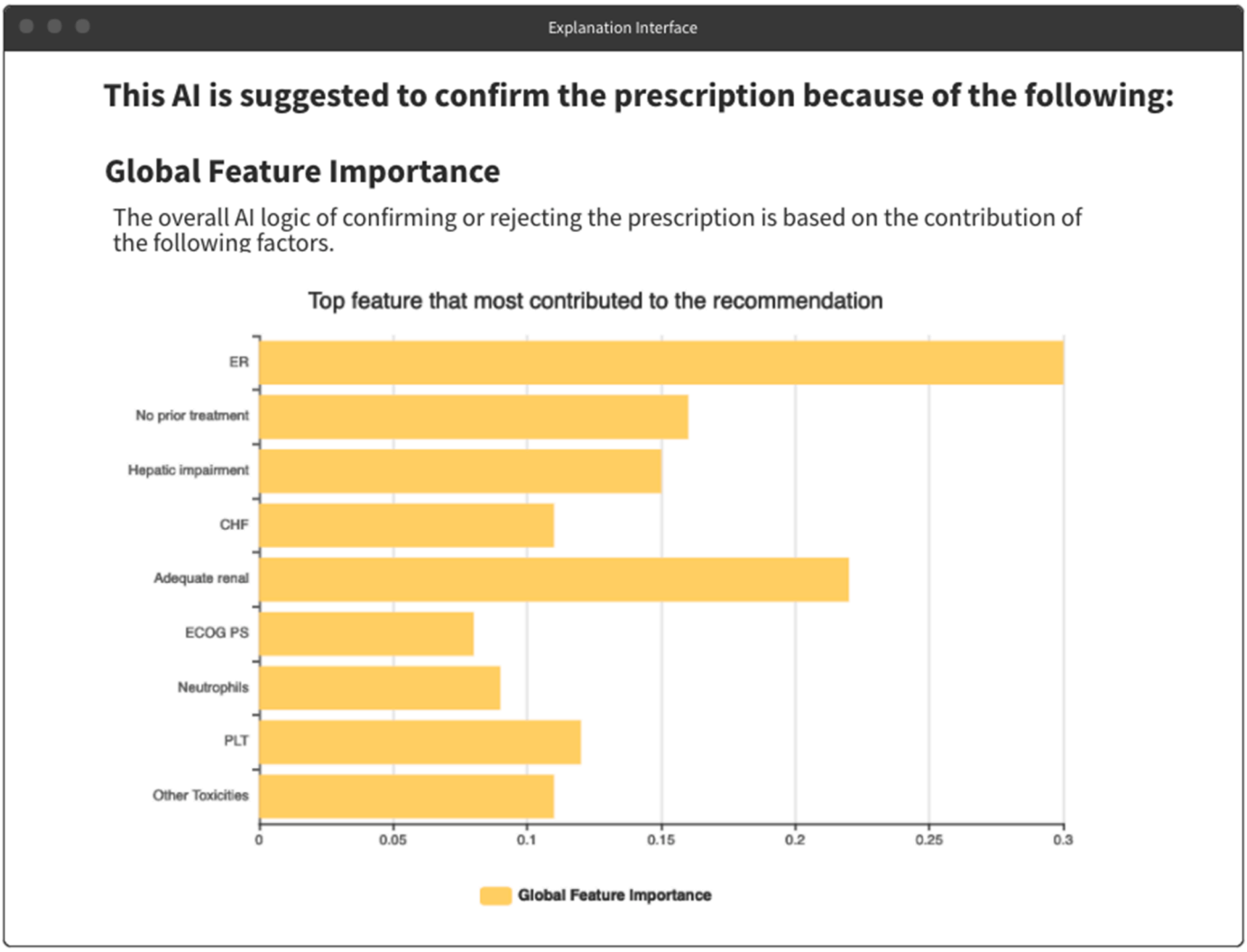

Fig. 8. Global Feature Importance explanation example.

that the development of such an explanation method is to help data scientists debug the ML model (Kaur et al., 2020; Ras et al., 2018), where the context of the task for such explanations in real-world scenarios is still a new area to discover. In psychology, such explanations with a lack of consistency and context are usually ignored by people (Keil, 2006). Further, recent research showed that explainability shall be designed and presented to support human decision-making strategy (Simkute et al., 2021). Overall, participants wanted explanations that are reflective of their task, i.e., task-centred explanations, for these explanations to be meaningful and reliable. In other words, explaining the logic of the algorithm shall be done in a way that is tightly coupled with the subject, i.e., the task for which the recommendation and the explanation are given. We recognise here that this can pose much more effort to systems engineers. Approaches to auto-generate and instantiate the algorithm-level explanation to a version that is also task-specific are still needed.

4.4. Assurances needs

Assurances in HCI literature are a design property that also applies to the intelligent tool so that they help users trust calibration (Israelsen and Ahmed, 2019). Assurances are indicators and performance metrics to indicate the actual capabilities of the intelligent tool. Interestingly, participants described assurances in terms of the XAI class. They also discussed how such assurances could support their intention to engage with AI explanations. Our data revealed two categories of assurances: *XAI class validity* and *XAI class capability*. Regarding the *XAI class validity*, participants described that they were unable to have guidance about trusting the provided explanation validity and correctness. Some suggested knowing the source of the data to assess the credibility of the explanation, with mentioning that it would also need to be up to date with the current changes in the task domain. P7 stated, "*As far as that is concerned, I cannot tell whether this explanation is right or wrong without knowing it is up to date*" and P13 added, "*from time to time we get emails to tell us the treatment x got recognised for diagnosing breast cancer. We need to ensure that the system knows this information*". Others also asked whether the explanation is generated based on training the AI on multiple data sources and references. P2 declared, "*reliable explanation should cover multiple medical sources and knowledge*". On the other hand, *XAI class capability* was related to the metrics used to evaluate the explanation in both AI and task domains. Participants argued that explanation verification with a medical expert should be performed according to accepted standards, incorporating best practices related to expert selection, elicitation protocols, bias avoidance, documentation and peer review. Moreover, participants raised questions that could be answered through clarity about the evaluation metrics used in the XAI models as introduced in a previous survey on interpretable machine learning models (Carvalho et al., 2019). Participants demanded information about the XAI class itself such as a) *accuracy* of an explanation on unseen cases, b) *fidelity* that shows how well the explanation is consistent with the

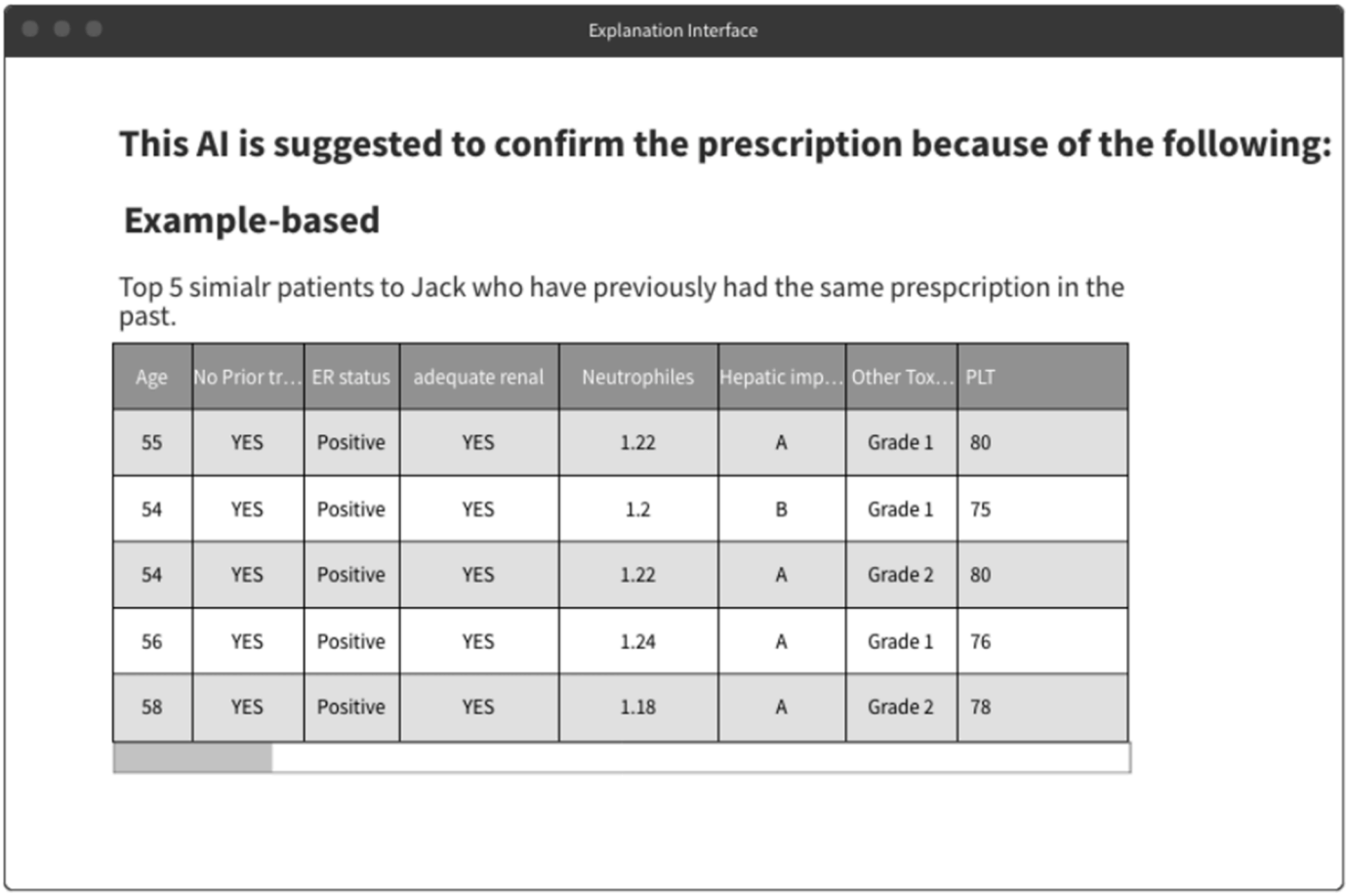

Fig. 9. Example-based explanation example.

underlying AI model, c) *stability* that represent how similar are the explanation for similar cases d) *representativeness* that describes how many cases could be covered by given explanation. For example, P1 mentioned, "*I was wondering if this explanation could be generalised for another patient*". We also argue that such assurances can be essential to repair user trust in an XAI class when the explanation generated from the XAI model is incomplete. Incomplete explanation refers to failure in the explainable model to generate complete meaningful information (Malhi et al., 2020).

4.5. Tailoring needs

Participants sought to tailor and customise the explanation output and its presentation to help them in contextualising and interpret the XAI explanation. This was also to meet their decision-making behaviour. For instance, participants were asked to set thresholds for similarity parameters in Example-based explanations. Participants identified Example-based explanations as a useful XAI class for their decision-making process and a way to calibrate their trust. P1 commented, "*I think this is crucial when I am sitting in the clinic and I need to make a decision, examples allow me to ask a whole range of questions even if it is one that what will your prognosis be what will the outcome be what how should I treat the patient how can I tell what events would be*". However, providing examples of similar cases confused participants in terms of the similarity definition, "*similarity is very hard to determine I am curious how the machine defines similarities*" [P5]. Similarity definition is a conflicting and complex problem in AI and XAI literature (Thrun, 2021). For this purpose, Wang et al. (2019) suggest several guidelines to support XAI developers to select context-based similarity methods based on how humans reason about explanations (Wang et al., 2019). For instance, explanations generated from distance-based methods, e.g., case-based reasoning (Aamodt and Plaza, 1994) and clustering models (Jain et al., 1999), are driven by inductive and analogical reasoning to understand why a certain case is considered similar or different. Our participants in this context wanted to control the explanation output by defining their own similarity metrics, "*I would like to ask for examples based on all the features in the system or subset features of the system find the similar patient for this recommendation*" [P14]. Another example of tailoring was encountered in Local and Global explanations when participants wanted to group a set of features to generate a group feature importance value. P8 described "*I think it was easier to read and recognise when this explanation [Local] groups patient history information in one value*". Rather than providing a static explanation, our findings suggest integrating a possibility to configure and tailor explanations and what they shall contain and how they are computed. This requirement could also be due to the need to fit the explanation into task workflow by finding, accessing, and focusing on intended information while minimising unrelated information (Petty and Cacioppo, 1986).

4.6. Multi-step explainability

It refers to users' explainability needs after utilising the main presented explanation. During our interviews, participants discussed that they had follow-up questions after reading AI explanations. They mentioned that such information would support them in validating the AI recommendation. For instance, participants figured out potential correlations between different features in the Global explanation in which they could not be able to validate their hypothesis, e.g., P2 commented, "*There is a lot of correlation between the treatment cycle and the patient history when you are aware that the system considering this*

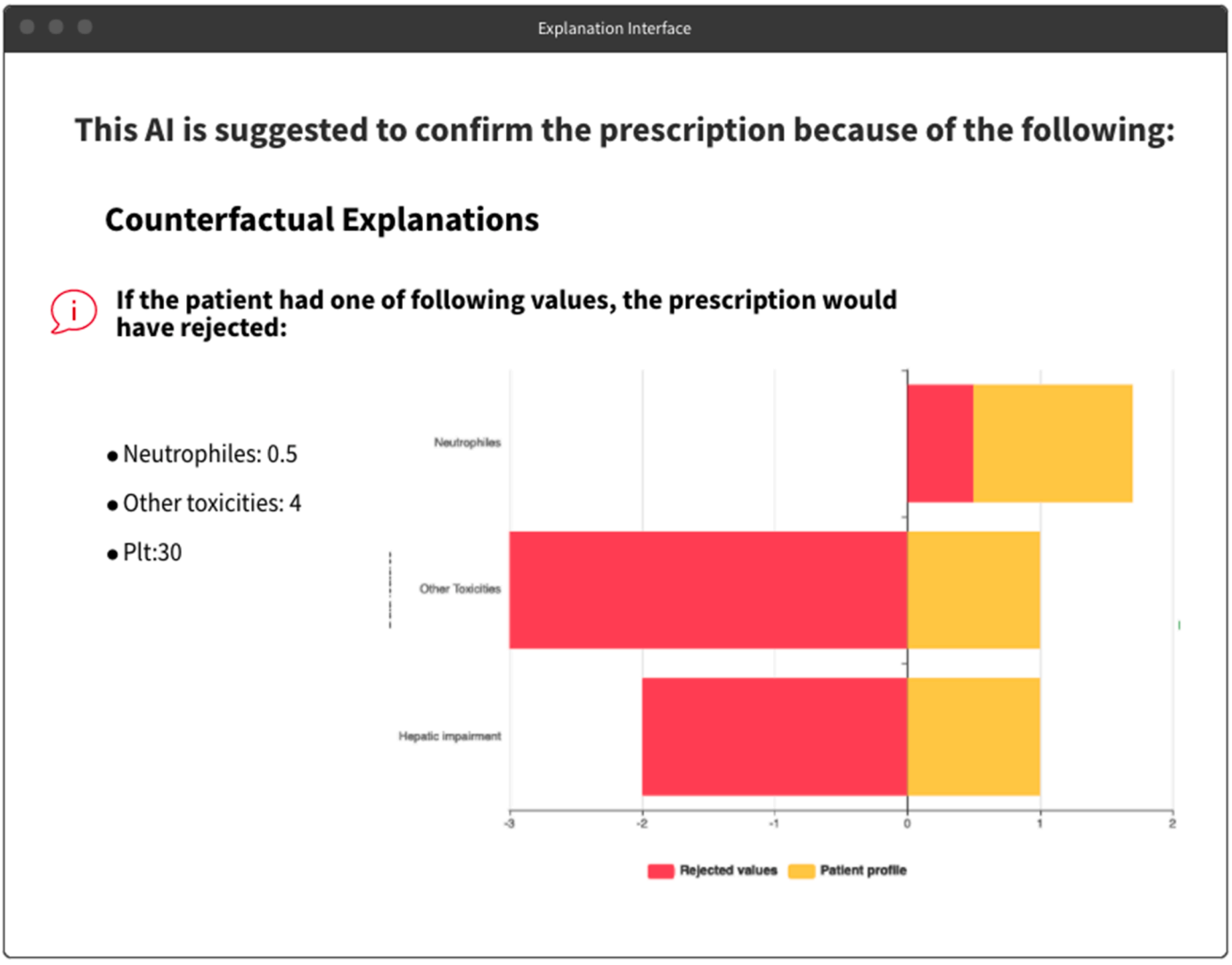

Fig. 10. Counterfactual exaplanation example.

correlation, I will be able to tell if the explanation is accurate". Another example encountered in Counterfactual explanations scenarios when participants wanted to explore the effect of a specific patient feature on the recommendation. P4 commented, "*for this patient scenario, I wanted to observe the AI decision if other toxicities have Grade 4*". A potential interpretation for a multi-step pattern in our data could be linked to previous learning and cognitive development literature (Kurkul and Corriveau, 2018). Humans are likely to ask follow-up questions when they did not receive a casual response to the explanation. This theory provides an interpretation of our participants' needs during the study for further information. Overall, explainability is a social and interactive process between the explainer and the explainee (Miller, 2019), and this may need to follow a multi-step interaction approach between the Human and the XAI interface. P3 commented in that context, "*I would have more accurate judgement if the AI … if a maybe I always can ask for an explanation about explanation*". As multi-step explainability interaction can provide more transparency about the XAI class and the AI model, it can support moving from explainability to causality. Causality is measured in terms of effectiveness, efficiency, and satisfaction related to the causal understanding of the human-understandable model (Holzinger et al., 2019). Developers of such explainable interfaces may need to collect data from end-users regarding the users' information needs after presenting the main explanations. Design considerations for the modalities of such multi-step explainability are also required to balance explainability and usability, e.g., chatbots.

5. Discussion

Consistent with previous research (Lundberg et al., 2018; Cai et al., 2019; Lai and Tan, 2019; Yang et al., 2020), we found that AI explanations with their different classes improved the overall performance of the Human-AI team. In this study, 41 medical practitioners performed 410 Human-AI tasks where these tasks were diverse in their XAI classes and the AI recommendation accuracy, i.e., we included correct and incorrect recommendations. While this work reports on results from a study in the medical domain, our findings are likely applicable to other contexts in collaborative Human-AI decision-making applications for expert users, primarily when high-stakes decisions are implemented. In this section, we provide a discussion of the potential implications of our results and design guidelines for enhancing trust calibration. We also discuss the limitations of our work.

5.1. Perceived understandability of an xai class can contribute to increasing or decreasing users' engagement with the xai interface

Across different XAI class conditions, XAI class was seen differently by our participants to affect their trust (increase or decrease) based on how they perceive understandability. Our results showed that Example-

based and Counterfactual explanations were more understandable to our participants than Local, Global, and No explanation conditions. The primary reason might be that these explanations are easy to understand by humans than local and Global explanations that require technical knowledge. According to psychological research, humans are more willing to engage with explanations when they are familiar, simple, and casually relevant (Colombo et al., 2017; Keil, 2006). This means that participants' willingness to engage with AI explanations may have correlated with how they perceived the understandability of the explanation during the study. Many participants discussed during the follow-up interviews that they skipped explanations when they could not interpret these explanations in their domain task. During the follow-up interviews, these observations were also confirmed when participants clearly suggested ways of making Local and Global explanations interpretable. For instance, participants discussed providing tools and modalities, e.g., interactive and dialogue explanations, to end-users to understand and contextualise explanations that they could not interpret. We argue that perceiving an explanation to be understood is crucial to increasing participants' engagement with AI explanations and therefore supporting appropriate trust judgement. A previous study by Cai et al. (2019) used an onboarding technique to guide users' understanding of the actual AI capabilities and limitations and ways of using it. They aimed to familiarise AI-based decision-making tool users with the AI and help users build appropriate trust. Our results extend their view and argue that users of XAI systems shall be familiarised with its explanations, e.g., usage scenarios or tutorials, to avoid potential misinterpretability and avoidance behaviour. Future work could explore how guiding users' understandability and interpretability of AI explanations could help calibrated trust. Also, approaches like Participatory Design (Schuler and Namioka, 1993) and Co-Design (Sanders and Stappers, 2008) that involve users early in the process will lead to more acceptable and interpretable explanations that fit their target groups.

5.2. Users' perceived reliability of the ai does not seem to be correlated with ai explainability

Importantly, we identified no significant change in the reliability of the AI between the baseline conditions and different XAI classes. The lack of difference in reliability scores suggests that participants' perceived reliability might not be related to showing explanations, which could be aligned with other AI components such as overall performance. For instance, Dietvorst et al. (2015) found that people are more likely to rely on AI when they can control the algorithmic output. The reliability dimension of trust could be related to another line of research that aims to increase trust with the AI not calibrate users' trust (Yin et al., 2019; Yu et al., 2019).

5.3. Human cognitive biases could have triggered overreliance during the study

Explanations are a common approach for supporting trust calibration in a Human-AI environment. Despite their benefits, recent studies showed that explanations could also be misused by participants (Naiseh et al., 2021c). Our research helps to unpack the complicated influence of explanation on behaviour, demonstrating how different XAI classes can affect human behaviour during Human-AI collaborative decision-making tasks. Consistent with prior research, explanations with their various classes (Example-based, Counterfactual, Local and Global explanation) improved the accuracy of the Human-AI task compared to No explanation conditions. However, our results showed that when the AI was not accurate and provided incorrect recommendations, explanations with its different classes did not help our participants recognise incorrect recommendations. Furthermore, our results showed that participants agreed more with the AI across all XAI classes than the No explanation scenario. These observations could be interpreted as participants over-relied on the AI when explanations were provided. Similar to Buçinca et al. (2021), we argue that the dual-process theory offers valuable insights to understand why the explanations may contribute to over-reliance. According to dual-process theory (Groves and Thompson, 1970), humans regularly operate on System 1 thinking, which follows heuristics and shortcuts when making decisions. The settings of our study were under an everyday human-AI collaborative decision-making task which might make our participants follow system 1. On the other hand, System 2 was infrequently triggered as it is slower and more effortful. System 1 might make our participants vulnerable to cognitive biases during the study, which results in their inability to recognise incorrect recommendations. These results are aligned with previous research (Buçinca et al., 2021; Naiseh et al., 2021c), which showed that designers of the XAI interface often assumed that users would engage cognitively with AI explanations and use them to calibrate their trust. Also, some studies showed that XAI users perceive explanations as a competency feature rather than applying analytical thinking to assess the AI output (Bansal et al., 2021). We argue that calibrating users' trust would require extra effort from both the XAI interface designers and XAI users. XAI designers to debiasing users' behaviour and XAI users to read and engage cognitively with AI explanations. An example of how such an XAI human-AI interface looks like has been introduced in recent publication (Holzinger and Muller, 2021). Finally, although we followed best practice to present explanations to end-users proposed in Laato et al. (2022), we also acknowledge that it is possible that trust calibration errors may have occurred due to the visual design that we used, specifically, the Local and Global explanations where participants perceived them to be less understandable. Nonetheless, our results highlight the importance of considering cognitive biases when designing the XAI interface for the trust calibration goal.

5.4. One explanation does not fit all users' needs during human-ai collaborative decision-making tasks

One explanation does not fit all users' needs during Human-AI collaborative decision-making tasks (Sokol and Flach, 2020). Our qualitative phase showed that users require XAI modalities and interaction techniques to help trust calibration. Participants viewed the XAI interface as a new interactive system that needs to be customisable to their needs and task requirements. An effective XAI interface needs to answer multiple users' questions and help users adjust the explanation accordingly. Participants posed several requirements while interacting with the XAI interface, such as tailoring, and multi-step explainability. This aligns with learning literature that shows that the learning process is personalised and is achieved via an explanatory dialogue (Jéirveléi'k, 2006). Following this dialogue process in XAI would make the interface engaging and acceptable to a wider range of users. Furthermore, allowing the user to customise explanations extends their utility beyond AI transparency (Sokol and Flach, 2020). For instance, the explainee can steer the explainability process to inspect errors, e.g., identify biases, and validate a hypothesis, e.g., for counterfactual explanations, users defined constraints on the number and type of features that may or may not appear in the explanation. We also note that the XAI literature may benefit from advances in visual analytics (VA). VA has been often used in providing interpretable ML models by underlying data understanding through an interactive visual interface (Kahng et al., 2017). Combining the techniques of VA with XAI algorithms would present a solution to 'one explanation does not fill all' problem. Finally, we argue that the case of calibrating users trust may require presenting multiple XAI classes in the XAI interface. Recent studies showed that different XAI classes could be useful to support various human reasoning methods and mitigate potential cognitive biases (Lim and Dey, 2010; Wang et al., 2019). For instance, Wang et al. (2019) discussed that counterfactual explanation is useful to mitigate anchoring bias which occurs when humans form a skewed perception and limit the possibility of exploring alternative options. Counterfactual explanations by its design determine what input features could change the AI recommendation and help

humans expose to alternative decisions and scenarios.

5.6. Limitations

We note several limitations to our work that warrant caution in generalising the results to other use cases or the field of XAI in general. The first limitation is related to the selected use case of the clinical decision support system. Although our use case is realistic and conducted with users who are experts in the task, participants' responses could introduce domain bias when evaluating XAI classes. Further, the study was conducted online, due to COVID-19 restrictions on social gatherings, and used hypothetical patient scenarios that focused on screening prescription as an ML classification problem. Our future work will focus on conducting the study settings in different use cases, e.g. human-swarm interaction and self-driving cars. Also, trust calibration is a process that is based on multi-criteria decision-making. In this paper, we study the outcome of this process through the decision made and self-reported measures of it. We do not study how trust is calibrated during the interaction between the AI and humans and this would require a different research design, perhaps based on non-intrusive measures, e.g., eye tracking can serve this purpose (Lu and Sarter, 2019). Furthermore, our study has not looked at the behaviour beyond one interaction with participants and more longitudinal studies are required to observe trust calibration in a long-term interaction.

Although our sample size met the requirements of a power analysis, a larger sample size for the quantitative phase would have made more conclusive results and enabled further analysis, e.g. linear regression. Another limitation of our work is that the sample was recruited from a mailing list containing three organisations. Those who volunteered to take part in the study may exhibit a different attitude than those who did not respond to our invitation in the sense of being interested in the subject. More research is needed to examine whether the different experiences of the participants in the task may have a possible effect on their requirements for an XAI class and impact their trust calibration. For example, novice users who are learning the task may be differently affected compared to expert users and may, similarly, have different requirements from the XAI interface. Finally, our qualitative data are only meant to raise questions for further investigation. We recommend further experimental research to quantitively evaluate solutions discussed in the qualitative section. For instance, we encourage experiments to examine whether multi-step explainability can indeed improve participants' overall performance during Human-AI collaborative decision-making tasks.

6. Conclusion

Explainablity is part of the overall knowledge discovery process and should be extended beyond the meaning of the discovered knowledge to cover meta-data about it including trust level. This is even more important when dealing with black-box algorithms and explaining the knowledge without revealing the underlying algorithms introduces a new set of challenges. In this paper, we studied the effect of four different explanation classes on trust calibration during Human-AI collaborative decision-making task. Our results indicated that Example-based and Counterfactual explanations were perceived as significantly understandable by participants in our experimental settings. On the other hand, interpreting Local and Global explanations required additional design considerations and interactive approaches to operationalise these explanations for end-users. Furthermore, our results showed that the presence of explanation with its different classes could introduce over-reliance on the AI, i.e., participants were more likely to follow AI recommendations when explanations were presented. These results pose future challenges for future work to explore XAI design modalities and principles to mitigate potential over-reliance risk when explanations are provided.

CRediT authorship contribution statement

Mohammad Naiseh: Conceptualization, Investigation, Methodology, Data curation, Formal analysis, Writing – original draft. **Dena Al-Thani:** Writing – review & editing. **Nan Jiang:** Supervision, Writing – review & editing. **Raian Ali:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

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.

Data availability

Data will be made available on request.

Acknowledgements

This work is partially funded by iQ HealthTech and the Bournemouth university PGR development fund. This work also acknowledges support from the Engineering and Physical Sciences Research Council (EP/V00784X/1).

Appendix A

Demographic and profile questionnaire

1 Please provide your age category.
   - ○ 20–30
   - ○ 30–40
   - ○ 40–50
   - ○ 50–60
2 Please provide your gender.
   - ○ Male
   - ○ Female
3 Approximately how long have you been practising clinically?
4 Please check all statements that apply regarding your level of experience screening chemotherapy prescriptions.
   - ○ I know what screening prescription is.
   - ○ I have used screening prescription software in practice.

5 Please indicate your level of agreement with the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Agree Strongly
Artificial Intelligence will play an important role in the future of medicine					
There are too many complexities and barriers in medicine for AI to help in clinical settings.					
I have reservations about using AI in clinical settings.					

Appendix B

We aim to provide explanatory information that helps medical practitioners to calibrate their trust in Collaborative Human-AI decision-making tools. We consulted with two AI experts and one medical expert, presenting them with the explainable interface and asked them for their expert opinion regarding the relevance of the explanations. We used these opinions as well as the results from our pilot study to refine the interface design. We presented ten individual patient scenarios to every participant. They have been initialised with fictional names and profiles to make them more realistic to our practitioners. Each scenario was accompanied by one different explanation class and was meant to be either a correct recommendation or an incorrect recommendation. We asked our participants to self-report their cognition-based trust components in each explanation class using 5 Likert Scale questions. Examples of mock-up interfaces are shown in Figs. 7– 10.

Perceived Reliability

R1 - The system always provides the advice I require to make my decision.
R2 - The system performs reliably.
R3 - The system responds the same way under the same conditions at different times.
R4 - I can rely on the system to function properly.
R5 - The system analyses problems consistently.

Perceived Technical Competence

T1 - The system uses appropriate methods to reach decisions.
T2 - The system has sound knowledge about this type of problem built into it.
T3 - The advice the system produces is as good as that which a highly competent person could produce.
T4 - The system correctly uses the information I enter.
T5 - The system makes use of all the knowledge and information available to it to produce its solution to the problem.

Perceived Understandability

U1 - I know what will happen the next time I use the system because I understand how it behaves.
U2 - I understand how the system will assist me with the decisions I have to make.
U3 - Although I may not know exactly how the system works, I know how to use it to make decisions about the problem.
U4 - It is easy to follow what the system does.
U5 - I recognize what I should do to get the advice I need from the system the next time I use it.

Appendix C

follow-up interview questions.

1 How would you summarise why the AI-supported decision tool made the recommendations?
2 What do you think about this explanation and how do you evaluate it in helping you to understand the AI recommendation?
3 How do you assess it in helping you to rely on the AI recommendation?
4 How do you assess it in helping you to identify the correctness of the AI recommendation?
5 Why might you be agreeing or disagree on [understandability, reliability and technical competence] of the recommendation and the explanation presented in this way?
6 What information led you to agree or disagree with the [understandability, reliability and technical competence] of the recommendation and its explanation?
7 What information is missing that might help you to assess the [understandability, reliability and technical competence] confidently or effectively?
8 What would you like to change about this explanation to help the assessment of the [understandability, reliability and technical competence]?

References

Aamodt, A., Plaza, E., 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications 7 (1), 39–59.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., 2019. Guidelines for human-AI interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems, pp. 1–13.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion 58, 82–115.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16.

Barrick, M.R., Mount, M.K., 1991. The big five personality dimensions and job performance: a meta-analysis. Personnel psychology 44 (1), 1–26.

Bayati, M., Braverman, M., Gillam, M., Mack, K.M., Ruiz, G., Smith, M.S., Horvitz, E., 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. PloS one 9 (10), e109264.

Birt, L., Scott, S., Cavers, D., Campbell, C., Walter, F., 2016. Member checking: a tool to enhance trustworthiness or merely a nod to validation? Qualitative health research 26 (13), 1802–1811.

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.

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. IEEE, pp. 160–169.

Cai, C.J., Jongejan, J., Holbrook, J., 2019. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th international conference on intelligent user interfaces, pp. 258–262.

Carroll, J.M., Olson, J.R., 1988. Mental models in human-computer interaction. Handbook of human-computer interaction 45–65.

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.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8 (8), 832.

Colombo, M., Bucher, L., Sprenger, J., 2017. Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. Frontiers in psychology 8, 1430.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., Cruz, F., 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. Artificial Intelligence 299, 103525.

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.

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

Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C., 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th international conference on intelligent user interfaces, pp. 275–285.

Ericsson, K.A., Simon, H.A., 1984. Protocol analysis: Verbal reports as data. the MIT Press.

Faulkner, S.L., Trotter, S.P., 2017. Theoretical saturation. The International encyclopedia of communication research methods 1–2.

Feng, S., Boyd-Graber, J., 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 229–239.

Flores, A.W., Bechtel, K., Lowenkamp, C.T., 2016. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation 80, 38.

Green, B., Chen, Y., 2019. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction 3 (CSCW), 1–24.

Groves, P.M., Thompson, R.F., 1970. Habituation: a dual-process theory. Psychological review 77 (5), 419.

Guesmi, M., Chatti, M.A., Vorgerd, L., Joarder, S.A., Ain, Q.U., Ngo, T., Zumor, S., Sun, Y., Ji, F. and Muslim, A., 2021. Input or Output: Effects of Explanation Focus on the Perception of Explainable Recommendation with Varying Level of Details. In IntRS@ RecSys (pp. 55-72).

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F. and Giannotti, F., 2018. Local rule-based explanations of black box decision systems. arXiv preprint arXiv: 1805.10820.

Hagras, H., 2018. Toward human-understandable, explainable AI. Computer 51 (9), 28–36.

Henelius, A., Puolamäki, K., Boström, H., Asker, L., Papapetrou, P., 2014. A peek into the black box: exploring classifiers by randomization. Data mining and knowledge discovery 28 (5), 1503–1529.

Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S.M., 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp. 1–13.

Holzinger, A., Muller, H., 2021. Toward human–AI interfaces to support explainability and causability in medical AI. Computer 54 (10), 78–86.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (4), e1312.

Holzinger, A., 2021. The next frontier: Ai we can really trust. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, pp. 427–440.

Israelsen, B.W., Ahmed, N.R., 2019. "Dave... I can assure you... that it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. ACM Computing Surveys (CSUR) 51 (6), 1–37.

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

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM computing surveys (CSUR) 31 (3), 264–323.

Jéirveléi'k, S., 2006. Personalised learning? New insights into fostering learning capacity. Schooling for Tomorrow Personalising Education, p.31.

Kahng, M., Andrews, P.Y., Kalro, A., Chau, D.H., 2017. A cti v is: Visual exploration of industry-scale deep neural network models. IEEE transactions on visualization and computer graphics 24 (1), 88–97.

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.

Keil, F.C., 2006. Explanation and understanding. Annual review of psychology 57, 227.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., Wong, W.K., 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE Symposium on visual languages and human centric computing. IEEE, pp. 3–10.

Kurkul, K.E., Corriveau, K.H., 2018. Question, explanation, follow-up: A mechanism for learning from others? Child Development 89 (1), 280–294.

Lai, V., Tan, C., 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 29–38.

Laato, S., Tiainen, M., Islam, A.N., Mäntymäki, M., 2022. How to explain AI systems to end users: a systematic literature review and research agenda. Internet Research 32 (7), 1–31.

Larasati, R., Liddo, A.D., Motta, E., 2020. The effect of explanation styles on user's trust. IUI 2020 Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies.

Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. Human factors 46 (1), 50–80.

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.

Lim, B.Y., Dey, A.K., 2010. Toolkit to support intelligibility in context-aware applications. In: Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 13–22.

Louthrenoo, W., Nilganuwong, S., Aksaranugraha, S., Asavatanabodee, P., Saengnipanthkul, S., Thai Study Group, 2007. The efficacy, safety and carry-over effect of diacerein in the treatment of painful knee osteoarthritis: a randomised, double-blind, NSAID-controlled study. Osteoarthritis and cartilage 15 (6), 605–614.

Lu, Y., Sarter, N., 2019. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Transactions on Human-Machine Systems 49 (6), 560–568.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., Lee, S.I., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering 2 (10), 749–760.

Madsen, M., Gregor, S., 2000. Measuring human-computer trust. In: 11th australasian conference on information systems, 53. Australasian Association for Information Systems, Brisbane, Australia, pp. 6–8.

Malhi, A., Knapic, S., Främling, K., 2020. Explainable agents for less bias in human-agent decision making. International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Springer, Cham, pp. 129–146.

Marshall, R.S., 2003. Building trust early: the influence of first and second order expectations on trust in international channels of distribution. International Business Review 12 (4), 421–443.

McAllister, D.J., 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of management journal 38 (1), 24–59.

MILLER, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267, 1–38.

NAH, F.F.-H., DAVIS, S., 2002. HCI research issues in e-commerce. Journal of Electronic Commerce Research 3, 98–113.

NAISEH, M., AL-THANI, D., JIANG, N., ALI, R., 2021a. Explainable recommendation: when design meets trust calibration. World Wide Web 24, 1857–1884.

NAISEH, M., CEMILOGLU, D., AL THANI, D., JIANG, N., ALI, R., 2021b. Explainable recommendations and calibrated trust: two systematic user errors. Computer 54, 28–37.

Naiseh, M., Al-Mansoori, R.S., Al-Thani, D., Jiang, N., Ali, R., 2021c. Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. In: 2021 8th International Conference on Behavioral and Social Computing (BESC). IEEE, pp. 1–5.

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

NARAYANAN, M., CHEN, E., HE, J., KIM, B., GERSHMAN, S., DOSHI-VELEZ, F., 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation arXiv preprint arXiv: 1802.00682.

NG, K.Y., CHUA, R.Y.J., 2006. Do I contribute more when I trust more? Differential effects of cognition-and affect-based trust. Management and Organization review 2, 43–66.

OATES, B.J., 2005. Researching information systems and computing. Sage.

Petty, R.E., Cacioppo, J.T., 1986. The elaboration likelihood model of persuasion. Communication and persuasion. Springer, New York, NY, pp. 1–24.

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. Springer.

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.

SANDERS, E.B.N., STAPPERS, P.J., 2008. Co-creation and the new landscapes of design. Co-design 4, 5–18.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R., 2021. Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE 109 (3), 247–278.

Schraagen, J.M., Elsasser, P., Fricke, H., Hof, M., Ragalmuto, F., 2020. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64. SAGE Publications, Los Angeles, CA, pp. 339–343. Sage CA:

SCHULER, D., NAMIOKA, A., 1993. Participatory design: Principles and practices. CRC Press.

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

SOKOL, K., FLACH, P., 2020. One explanation does not fit all. KI-Künstliche Intelligenz 1–16.

Sokol, K., Flach, P.A., 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. SafeAI@ AAAI.

THRUN, M.C., 2021. The Exploitation of Distance Distributions for Clustering. International Journal of Computational Intelligence and Applications 20, 2150016.

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

Wang, N., Pynadath, D.V., Hill, S.G., 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp. 109–116.

Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, M.R., Tai, Y.W., 2020. Towards global explanations of convolutional neural networks with concept attribution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8652–8661.

Yang, F., Huang, Z., Scholtz, J., Arendt, D.L., 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 189–201.

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.

Yu, K., Berkovsky, S., Taib, R., Zhou, J., Chen, F., 2019. Do i trust my machine teammate? an investigation from perception to decision. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 460–468.

Zhang, Y., Liao, Q.V., Bellamy, R.K., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305.