Journal of Responsible Technology 17 (2024) 100076

Contents lists available at ScienceDirect

Journal of Responsible Technology

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

C-XAI: A conceptual framework for designing XAI tools that support trust calibration

Mohammad Naiseh [a,*], Auste Simkute [b], Baraa Zieni [c], Nan Jiang [a], Raian Ali [d]

[a] Faculty of Science and Technology, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK
[b] The University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB UK
[c] Faculty of Computer Science and Engineering, University Carlos III de Madrid, Av. de la Universidad, 30, 28911 Leganés, Madrid, Spain
[d] College of Science and Engineering, Hamad Bin Khalifa University, Qatar

ARTICLE INFO

Keywords:
Explainable ai
Human-centred design
Participatory design
Human-AI teaming

ABSTRACT

Recent advancements in machine learning have spurred an increased integration of AI in critical sectors such as healthcare and criminal justice. The ethical and legal concerns surrounding fully autonomous AI highlight the importance of combining human oversight with AI to elevate decision-making quality. However, trust calibration errors in human-AI collaboration, encompassing instances of over-trust or under-trust in AI recommendations, pose challenges to overall performance. Addressing trust calibration in the design process is essential, and eXplainable AI (XAI) emerges as a valuable tool by providing transparent AI explanations. This paper introduces Calibrated-XAI (C-XAI), a participatory design framework specifically crafted to tackle both technical and human factors in the creation of XAI interfaces geared towards trust calibration in Human-AI collaboration. The primary objective of the C-XAI framework is to assist designers of XAI interfaces in minimising trust calibration errors at the design level. This is achieved through the adoption of a participatory design approach, which includes providing templates, guidance, and involving diverse stakeholders in the design process. The efficacy of C-XAI is evaluated through a two-stage evaluation study, demonstrating its potential to aid designers in constructing user interfaces with trust calibration in mind. Through this work, we aspire to offer systematic guidance to practitioners, fostering a responsible approach to eXplainable AI at the user interface level.

1. Introduction

Artificial Intelligence (AI) is increasingly being used to support human decision-making in high-stakes scenarios such as the healthcare (Yu, Beam & Kohane, 2018), defence (Clark et al., 2022), and finance sectors (Cao, 2022). Full automation is often undesirable due to ethical and legal concerns (Naiseh, Jiang, Ma & Ali, 2020), as well as the importance of the outcome (Naiseh et al., 2020) . Instead, combining human knowledge and machine intelligence is expected to improve decision-making quality, whether made by humans or AI (M. Naiseh, Al-Thani, Jiang & Ali, 2021). In this paper, we refer to these tools as human-AI decision-making tools. To be successful, such tools should be designed to support human decision-makers in forming a correct mental model of AI capabilities and limitations (Zhang, Liao & Bellamy, 2020). This allows human decision-makers to judge when to trust or distrust AI recommendations. Failure to calibrate trust can lead to degraded performance of the human-AI team and costly errors in high-stakes application scenarios (Naiseh, Al-Thani, Jiang & Ali, 2023). The research community has discussed the challenges of understanding and developing accurate mental models of AI, particularly as opaque ML black-box models are increasingly used. For example, research has shown that humans may fail to understand certain outputs due to the dynamic and opaque nature of ML algorithms (Bansal et al., 2019). This can lead to over-trusting incorrect recommendations or under-trusting correct ones (Bussone, Stumpf & O'Sullivan, 2015; Naiseh et al., 2021). A design goal that aims to attain and manage trust refers to calibrated trust (Naiseh et al., 2021).

Studies have shown that humans require a user interface to help calibrate their trust in AI (Bansal et al., 2019; Naiseh et al., 2021; Zhang et al., 2020). This interface should reflect the logic of the AI and provide a rationale for its recommendation. By explaining the AI output, humans can decide whether to follow or reject the AI recommendation and maintain an appropriate level of trust. The benefits of generating explanations from AI models have gained interest in an emerging field

* Corresponding author.
E-mail address: mnaiseh1@bournemouth.ac.uk (M. Naiseh).

https://doi.org/10.1016/j.jrt.2024.100076

Available online 26 January 2024

called eXplainable AI (XAI) from various disciplines, including Psychology (Taylor & Taylor, 2021), Human-Computer Interaction (Naiseh et al., 2021), Social Sciences (Miller, 2019), and Law (Hacker, Krestel, Grundmann & Naumann, 2020). The assumption behind XAI to support trust calibration and improve Human-AI performance has its theoretical foundation since humans will interact with AI explanations and decide whether the rationale behind the AI decision can be trustworthy (Miller, 2019; Taylor & Taylor, 2021) . However, recent empirical studies have shown that communicating explanations to human decision-makers do not always result in improved trust calibration (Bussone et al., 2015; Naiseh et al., 2023; Naiseh, Cemiloglu, Al Thani, Jiang & Ali, 2021). People working with AI still make trust calibration mistakes by following incorrect recommendations or rejecting correct ones. This failure of explainable systems to enhance trust calibration has been linked to several human factors, such as cognitive biases (Naiseh et al., 2021), human laziness (Wagner & Robinette, 2021), and a lack of curiosity (Hoffman, Mueller, Klein & Litman, 2018). Overall, users of explainable systems fail, on average, to calibrate their trust, meaning that human decision-makers working with an AI can still follow incorrect recommendations or reject correct ones.

Another factor that could contribute to failure in trust calibration has a technical dimension (Sokol & Flach, 2020). Imagine a scenario in which a team is conducting explainability requirement analysis for a cancer detection Human-AI task. The team has determined that the Local Feature Importance explanation can meet the task requirements and enable humans to understand the decision's rationale. At this stage, the team may face a difficult decision during the development stage. Which Local Feature Importance method can provide better Human-AI performance? What is the impact of a specific method on trust calibration? Recent studies have revealed that the increasing availability of XAI methods presents challenges for explainable systems designers in selecting between available methods (Naiseh et al., 2023; Sokol & Flach, 2020). It is even more difficult to keep track of these methods and understand their effect on trust calibration and Human-AI performance in general (Sokol & Flach, 2020).

While various methodological approaches exist to support designers in crafting effective eXplainable AI (XAI) interfaces, notable frameworks address specific aspects of the design process. Eiband et al. (Eiband et al., 2018) propose a framework that integrates transparency design practices into development, prioritising user understanding and control over AI-driven decisions. However, their approach lacks consideration for the technical capabilities of XAI methods. In contrast, Sokol and Flach (Sokol & Flach, 2020) introduce the "Explainability Fact Sheets," a novel framework systematically evaluating and comparing XAI methods across key technical dimensions like functionality, operational aspects, usability, safety, and validation. Although comprehensive in assessing technical aspects, this framework overlooks the potential impact of technical properties on user behaviour. This paper contends that empowering designers to understand how technical properties influence user behaviour is crucial for making informed decisions on trust calibration and Human-AI performance. For instance, in a specific case, the design team opted for LIME (Ribeiro, Singh & Guestrin, 2016), an XAI method with low generalizability. This implies that explanations generated by the LIME algorithm cannot be generalised beyond specific AI recommendations. Users relying on LIME explanations may risk forming inaccurate interpretations and generalising explanations, contributing to trust calibration errors. Addressing such technical limitations, alongside human factors, becomes imperative when designing XAI interfaces with a trust calibration objective.

In this paper, we emphasise on the same argument in Naiseh et al. (2023); Zhang et al. (2020)) that designing for trust calibration is different from designing to inspire trust in AI. Inspiring trust can be achieved at a global level and may not necessarily require humans to comprehend the AI. For instance, providing metrics for AI's overall performance has been found to increase trust in AI, but it does not calibrate trust (Lai & Tan, 2019). In contrast, designing for the calibrated trust should focus on the recommendation level, where humans can scrutinise each AI recommendation and determine whether it is correct or incorrect (Naiseh, Al-Mansoori, Al-Thani, Jiang & Ali, 2021). Our argument is that achieving calibrated trust as a design goal may require additional effort from human decision-makers. For example, they may need to engage with AI explanations to understand the AI's reasoning. Moreover, designing for trust calibration entails equipping the interface with interactive features that encourage decision-makers to adopt desirable behaviour and address the technical properties of XAI methods during the design process (Naiseh et al., 2021; Simkute, Surana, Luger, Evans & Jones, 2022).

We argue that operationalizing XAI methods at the user interface level requires a systematic approach that addresses both technical and human factors. This paper presents a design method Calibrate trust in eXplainable AI (C-XAI), which is tailored to help trust calibration in XAI interface design. The method addresses technical and human factor challenges by identifying technical properties of XAI algorithms that may introduce trust calibration errors and helps produce designs to mitigate these errors. Trust calibration risk is defined as a limitation in the interface design that may contribute to trust calibration errors or does not prevent trust calibration errors. C-XAI was evaluated by multidisciplinary experts and end-users, and the results showed that the method helped stakeholders understand the design problem and develop XAI designs with trust calibration problem in mind. The evaluation investigated the effectiveness, completeness, clarity, engagement, and communication between different stakeholders.

2. C-XAI framework

C-XAI framework follows a specific structure to support the design team consisting of professionals from various disciplines, including system analysts, AI experts, and psychologists, throughout the design process. C-XAI adopts a participatory design approach, ensuring active involvement of all relevant stakeholders in the initial stages of XAI interface design. Its objective is to offer organisations a systematic approach to developing user-centric XAI interfaces for their human-AI systems, with the aim of minimising potential trust calibration errors as highlighted in the introduction section. Participatory design approach can be effective in designing XAI interfaces by ensuring user-centred design, incorporating diverse perspectives, fostering contextual relevance, addressing ethical considerations, and facilitating iterative design based on real-world testing, ultimately enhancing trust, transparency, and usability in XAI systems. This user-involved approach helps bridge technical aspects with user expectations and contributing to the successful adoption of XAI technologies specially in the context of Trust and Trust calibration given the human-centric nature of it.

The C-XAI framework consists of four main phases: Identification, Assessment, Selection and Implementation, and Evaluation (refer to Fig. 1). Each phase encompasses several activities (refer to Table 1). To support users in completing each activity, C-XAI framework provides additional documents. All relevant documents and sheets for C-XAI v3 can be accessed at [https://bit.ly/3gmdPpy]. The upcoming sections will provide a detailed explanation of the phases and activities involved in C-XAI v3.

The first phase of C-XAI is the identification phase, where representative users are recruited to gather requirements for explanations pertaining to a human-AI task. Subsequently, the assessment phase evaluates both technical and human factors of these explanation requirements, focusing on identifying whether the explanations could potentially trigger trust calibration errors. The outcomes of the assessment enable the design team and system analysts to gain a deeper understanding of the design problem and propose suitable design solutions in the Selection and Implementation phase. Throughout this phase, the multidisciplinary team of experts engages in various activities to mitigate potential risks associated with the C-XAI process.

Finally, the iterative evaluation stage involves subsequent meetings

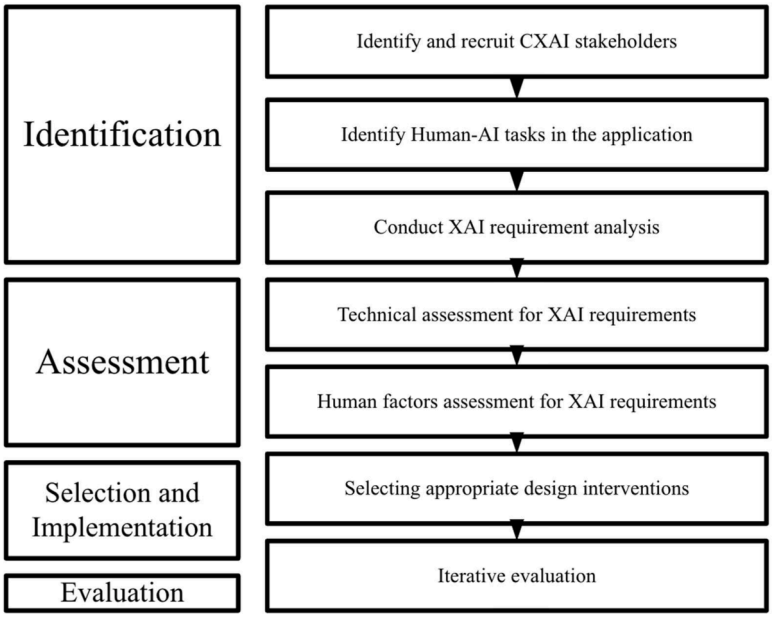

Fig. 1. Phases and activities of C-XAI.

Table 1
Description of C-XAI activities.

Process building blocks	Description
Identify and recruit CXAI stakeholders	This activity involves the process of identifying and reaching out to representative users of the Human-AI tool as well as engaging relevant stakeholders in the design process.
Identify Human-AI tasks in the application	This refers to the task analysis activity conducted on the application, aiming to identify specific tasks in which AI recommendations are provided to humans.
Conduct XAI requirement analysis	This pertains to the elicitation process aimed at determining the explanation requirements for each task, taking into account the users' needs. This process involves identifying suitable XAI algorithms and models to generate explanations that are relevant to the tasks and users.
Technical assessment for XAI requirements	This refers to a systematic technical evaluation of XAI models and algorithms relevant to the human-AI task.
Human factors assessment for XAI requirements	This refers to a systematic evaluation of various human factors, specifically focusing on the risks associated with trust calibration, for XAI models and algorithms.
Selecting appropriate design interventions	This activity promotes the generation of innovative solutions by designers to address the risks identified in the assessment phase or advises the AI team to consider utilizing alternative XAI methods.
Iterative evaluation	As each design is implemented, it must be evaluated in the context of the task. This might lead to iterating through the selection and evaluation phases.

to re-evaluate the design and assess potential trust calibration risks. This section introduces the third version of C-XAI v3, which has been developed after two rounds of evaluations described in Section 3.

2.1. Phase 1: identification process

The first stage of the C-XAI framework is the identification phase, which aims to identify the XAI methods needed for the Human-AI task. This phase involves three key activities. The first activity is to invite relevant stakeholders and representative users to participate in the design process. The second activity is a task analysis conducted by the system analyst to identify the Human-AI tasks in the application. The third and final activity is a requirements analysis to identify potential XAI methods for each Human-AI task. The following sections will provide more detail on each of these activities.

2.1.1. Activity 1: identify and recruit C–XAI stakeholders

The stakeholder identification activity aims to involve relevant stakeholders who can contribute to the success of the C-XAI framework. These stakeholders will ensure that the XAI interface can assist users in trust calibration and minimise trust calibration risks. Before starting the design activities, the system analyst should identify and invite the stakeholders who will participate in the design process. Table 2 provides a suggested list of stakeholders, with some directly involved in the entire design method and others participating in specific phases.

Initially, the system analyst is tasked with inviting representative users and relevant stakeholders to participate in the design process. To enhance the validity and credibility of the collected requirements, the system analyst is recommended to seek diversity within the sample. Previous studies have demonstrated that factors like the level of AI knowledge (Liao, Gruen & Miller, 2020), domain knowledge (Sokol & Flach, 2020), and curiosity level (Hoffman et al., 2018) can contribute to enriching the explainability requirements. Various recruitment methods, including convenience samples through organisational mailing lists, can be employed for this purpose.

Once representative users and stakeholders have been identified, an induction session can be organised to acquaint them with different C-XAI activities. During this session, the system analyst can utilise C-XAI supporting documents, templates, and materials to familiarise C-XAI stakeholders with the activities. It is advised for the system analyst to clarify that certain templates necessitate discussions amongst stakeholders, while others can be completed independently without the analyst's intervention. The induction session serves to stimulate stakeholders' thinking and prepare them for the upcoming activities. The system analyst can choose to conduct a separate short focus group or incorporate the induction session at the beginning of each C-XAI activity, along with the relevant documents, for instance, around 30 min prior to the activity.

2.1.2. Activity 2: identify human-AI tasks in the application

In this activity, the system analyst is responsible for identifying the various tasks within the application that involve issuing AI

Table 2
C-XAI framework Stakeholders.

Stakeholder	Description	Degree of participation
Representative users	This refers to the users who will utilize the human-AI decision-making tool.	Representative users are expected to involve in the identification and evaluation phases.
System Analyst	This pertains to the individual responsible for guiding C-XAI activities and gathering the collected requirements. The system analyst is expected to possess expertise in software engineering, along with sufficient experience in AI, human factors, and usability.	The role of the system analyst is to provide guidance throughout the design process and actively participate in all phases of C-XAI.
AI experts	This refers to the individual responsible for conducting technical assessments of XAI methods for each Human-AI task.	AI expert is expected to be involved in the identification and assessment phases.
Design team	This refers to a team of individuals who are responsible for designing the XAI interface.	They are involved in the assessment, implementation, and evaluation phases.
Domain expert	It refers to the people who have experience in the Human-AI task, e.g., an experienced doctor for a cancer detection task.	They are responsible for analysing and evaluating potential constraints and requirements for explanations in the Human-AI task.
Psychologist	This refers to individuals who possess psychological knowledge and a relevant background in cognitive biases, human behaviour, and decision-making theories. They play a crucial role in assisting the design team in understanding the psychological state of users and their decision-making strategies.	They are fully involved in all the stages of the design method, starting from the identification process to evaluating the generated XAI interfaces.

recommendations to human decision-makers. To facilitate this process, C-XAI recommends utilising a task analysis sheet called the Human-AI Task Analysis (HAI-TA), as depicted in Fig. 2. This sheet builds upon traditional task analysis techniques, such as those proposed by Annett (Boehm & DeMarco, 1997) and Crandall et al. (Harding, 1998) and has been expanded to support the requirement analysis of explainable AI. The completion of this activity is marked by filling out the task column in the HAI-TA sheet.

2.1.3. Activity 3: conduct XAI requirement analysis

Several studies have highlighted that humans' requirements for XAI in a human-AI task differ based on the specific task and the target audience (Naiseh et al., 2023; Naiseh et al., 2021, 2021). Consequently, there is no one-size-fits-all solution regarding the types of explanations to offer users during a human-AI task. In Activity 3, the analyst's role is to identify the XAI needs for each task outlined in the HAI-TA Sheet. This enables the system analyst and AI experts to select XAI methods that align with the users' requirements. The outcome of this stage is a compilation of explanation types and XAI methods that are tailored to meet the users' needs.

The activity consists of two steps: (1) eliciting the XAI needs of decision-makers and (2) translating these needs into available XAI methods. The outcome of this activity is a comprehensive list of XAI requirements for each human-AI task, along with potential XAI methods and techniques that can be utilised to generate these explanations. To support system analysts in completing this activity, we have developed A3D1 (A refers to the Activity number and D refers to the. Document ID) , which contains a range of suggested elicitation methods proposed by Hoffman et al. (Hoffman et al., 2018). Additionally, it provides the system analyst with the advantages and disadvantages of each method to facilitate informed decision-making. Fig. 3 displays only two rows of the A3D1 to conserve space in this paper and full version can be accessed in Appendix A – Table A2. C-XAI also offers SD2 (SD refers to a Supportive Document) that outlines best practices during the XAI requirements analysis (Appendix A – Table A1). This document is built upon earlier research conducted in the domain Naiseh et al. (2020); Naiseh et al. (2021), 2023; Naiseh et al. (2021), 2021). After completing this activity, system analysts shall complete the HAI-TA analysis sheet with relevant information related to XAI requirements and potential additional needs.

Once the XAI requirements have been identified, the AI expert is responsible for translating those needs into the currently available XAI methods. To aid in this process, C-XAI offers A3D2 (Fig. 4), which is a document that links potential user requirements and questions with existing XAI methods. This document is based on our previous systematic literature review of XAI methods (Naiseh et al., 2021), and it is supported by earlier work by Liao et al. (Liao et al., 2020). AI experts can refer to this document as they translate users' requirements into available XAI methods. Upon concluding this activity, the system analyst and AI experts should fill out the HAI-TA analysis sheet with pertinent information concerning XAI methods.

2.2. Phase 2: assessment

The objective of the assessment phase is to assist the design team in assessing the technical properties of each XAI method and evaluating whether specific technical properties may require further consideration at the user interface level. The input for this phase is a fully completed HAI-TA sheet obtained from the identification phase. The assessment phase consists of two primary activities: (i) technical assessment of each XAI method, and (ii) human factors assessment for each XAI method. We provide further details on each activity in the following sections.

Task	**XAI requirements**	**XAI method**	**Additional needs**
Prescription classification	Local Explanations	LIME	Data features correlation Data features annotation Data features grouping Explain the terminology of the AI when needed

Fig. 2. HAI-TA analysis sheet for prescription classification AI tool used later in the evaluation sectio.n

A3D1	**SHEET TITLE:** Methods to elicit XAI requirements
1	**Method:** Think-aloud problem solving, where participants think-aloud during a decision-making task. ***Strengths:*** It offers rich information about the users' mental model. ***Weaknesses:*** The process of transcription and data analysis might be time consuming.
2	**Method:** Think-Aloud protocol with specific question and answering activities. ***Strengths:*** It enables the researcher to target specific issues during the decision-making task. ***Weaknesses:*** It might introduce bias as it depends on the researcher skills in designing the study.

Fig. 3. SD2 is a supportive document offered during C-XAI design process to help the system analyst choose between different methods to elicit XAI requirements for human-AI task.

A3D2	TITLE: Translating requirements needs into XAI methods		
Main class	**Sub-class**	**Question – Content – Scope (QCS)**	**XAI methods examples**
Global explanations	**Global feature importance,**	*Ranking the data features* Why – Importance - Model	*PFI, RFE, Feature Importance using Tree-based Models, LASSO Regression, PCA*
		Correlation between features Why – Dependences – Model	*CFS, PLS, Mutual Information, Distance Correlation, CCA*
	Decision tree approximations	*Trace the model output* Why – trace – Model Why not – trace – Model How – trace – Model What if – trace - Model	*Interpretability via model extraction, There-Rule Extraction*
	Rule extraction	*AND-OR rules* How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model *IF-ELSE rules* How – trace – Model What if – trace - Model Why – trace – Model Why not – trace – Model	*Column Generation*
Local Explanations	**Local feature importance**	Why – importance – recommendation	*LIME, SHAP, PDP, ICE*
	Local Trees	How – trace – recommendation What if – trace - recommendation Why – trace – recommendation Why not – trace – a recommendation	*G-REX, Anchors*
Example-based	**Prototype**	When – exemplar – recommendation What else – exemplar – recommendation	*Prototype Selection*
	Counterfactual	When – exemplar with small changes – recommendation What else - exemplar with small changes – recommendation	*Inverse Classification, CEM*
	Influential	When – abstract exemplar – recommendation	*Harnessing Adversarial*
		What else – abstract exemplar – recommendation	
Counterfactual	**Feature Influence**	What if – Influence - recommendation	*Individual Conditional Expectation*
	Counterfactual features	When – influence – recommendation Why – influence – recommendation How to – influence - recommendation	*DiCE, ACE, LICE, CEIP*
Confidence		How accurate – uncertainty - recommendation How accurate is – uncertainty – model	*Probability-based methods, Drop-out based methods*

Fig. 4. A3D2 is a supportive document developed by C-XAI to assist AI experts in translating users' needs into available XAI methods. The "Main Class" and "Sub-class" columns refer to families of XAI methods based on their output information. The document classifies XAI methods based on three main pieces of information: 1. "Question" specifies the type of user questions that can be answered by an explanation generated from a particular sub-class. 2. "Content" describes the type of information generated by a sub-class. 3. "Scope" provides information on whether the explanation is at the model or recommendation level. The document also provides examples of XAI methods in the final column.

2.2.1. Activity 4: technical assessment for XAI requirements

XAI methods hold a diverse range of technical properties and features that may require attention at the XAI interface design level (Sokol & Flach, 2020). Neglecting such properties during the XAI interface development could trigger trust calibration errors (Naiseh et al., 2023; Naiseh et al., 2021). Our argument in this paper is that following a systematic assessment of the XAI method would help anticipate potential trust calibration risks. For instance, when considering that users might develop habits with the XAI interface, i.e., people become gradually less interested in the details of the explanation and overlook and perceive it to be familiar to them. This risk is critically important with XAI methods that have a high novelty as technical property, i.e., a high probability of generating new information for users each time. Good design practice in such cases is to highlight the new information in the XAI interface to guide users' attention and challenge users' habits. For this purpose, C-XAI framework provides technical assessment sheets for XAI methods based on Explainability Facts Sheets (EFS) framework (Sokol & Flach, 2020). EFS has five main dimensions to assess the technical properties of XAI methods:

- Functional. This dimension can help to assess whether a particular XAI method is suitable for the underlying AI algorithm and the human-AI task. For instance, whether the XAI method can be applicable to type of machine learning problem: classification (binary, multi-class or multi-label), regression or clustering.
- Operational. This dimension can support designers in understanding how users can interact with an explanation. For example, whether the explanation generated by given XAI method can be static or interactive.
- Usability. It helps to evaluate the XAI method based on theories of explainability in social sciences. For instance, soundness property refers to the degree of accuracy, reliability, and correctness in the explanations provided by the XAI models. A sound explanation should faithfully represent the underlying logic and decision-making processes of the AI model, ensuring that it aligns with the actual functioning of the system.
- Safety. Explanation models communicate partial information about the data set used to train the AI-based system. The safety dimension evaluates the effect of the XAI method on the security and privacy of AI systems.
- Validation. It ensures explanations are accurate and effective. It checks consistency with known truths, analyses user responses through studies, uses objective metrics to assess quality, and allows for continuous improvement based on evaluation findings.

The system analyst in a collaboration with AI experts is expected to complete A4-D1.1 – D1.5. We provide an example of usability

A4D1.3 Usability Requirements Assessment

TASK ID: ……. XAI method: ……...

How **truthful** is the XAI method with the underlying ML model?
○ Low ○ Medium ○ High
How well can the generated explanation be **generalised** beyond a particular recommendation?
○ Low ○ Medium ○ High
Describe any **contextual information** that the explanation shall accompany; This could be any information needed by the users to interpret the explanation correctly.
Does the XAI method allow **interactive** interaction?
○ Yes ○ No
Does the explanation have an **actionable** nature?
○ Yes ○ No
Does the explanation inherent **time order** of the event?
○ Yes ○ No
What is the degree of **coherence** within the XAI method?
○ Low ○ Medium ○ High
What is the degree of XAI method **novelty**?
○ Low ○ Medium ○ High
What is the degree of **complexity** inherited in the generated explanation?
○ Low ○ Medium ○ High
Are there any other assumptions or operational aspects related to the explanation method?

Fig. 5. Assessment Sheet for usability properties provided by C-XAI.

assessment document in Fig. 5. The assessment has five main dimensions:

2.2.2. Activity 5: human factors assessment for XAI requirements

After assessing the XAI method(s) in Activity 4, Activity 5 aims to help the design team identify properties of XAI method(s) that may pose a risk to trust calibration. Trust calibration risk assessment involves assessing the significance and acceptability of risk probabilities and consequences to trust calibration. This approach has been found to be effective in identifying design requirements during requirement elicitation processes (Boehm & DeMarco, 1997). This stage requires collaboration between system analysts, domain experts, AI experts, and psychologists. Conducting risk assessment would enable the analyst and design team to better understand the XAI method's applicability in the Human-AI task and anticipate potential trust calibration risks. This assessment activity takes the technical assessment documents completed in Activity 4 to evoke brainstorming on potential risks based on the technical properties of XAI method. Trust calibration risk assessment is not intended to be a prediction of the users' behaviour or system functionalities, but it helps to identify and assess potential risks associated with technical properties of XAI method(s). C-XAI framework provides five risk assessment templates (one for each technical assessment dimension). We designed these templates based on risk management guidelines outlined by Harding (Harding, 1998) – See Fig. 6. The proposed risks in the templates are based on our earlier work (Naiseh et al., 2021). In this stage, the system analysts shall complete A5D1.1 – A5D1.5 in collaboration with domain experts, psychologists and designers. The C-XAI framework also provides several examples of mapping between XAI method's technical properties and potential risks.

2.3. The selection and implementation phase

The selection and implementation phase takes risk assessment sheets from the assessment phase and determines mechanisms that are capable of mitigating potential trust calibration risks. As there are almost multiple ways to address each trust calibration risk. This is a creative process that will likely depend on the design team; however, C-XAI framework presents several design principles and guidelines for designing XAI interfaces to help trust calibration. To assist this process, C-XAI provides A6D1 which highlights trust calibration design principles that maps potential trust calibration risks and design solutions to help mitigate those risks. A6D1 is based on the results from previous work that was

A5D1.3	**Title :Trust Calibration Risk assessment – Usability Dimension**
TASK ID: ……	Explanation method: ……..
Provide a title for this assessment - This section depicts what trust calibration risk is.	
Risk identification. Based on the Usability dimension assessment sheet, what kind of risks can be identified?	
Please identify which XAI usability properties might introduce risks to the current Human-AI task? ○ Truthfulness ○ Completeness ○ Contextfullness ○ Interactivity ○ Actionability ○ Novelty ○ Time ordering ○ Coherence ○ Complexity ○ Other: ----------	
Risk analysis. Based on the selected properties, what kind of risks can be identified?	
Skipping. ○ Lack of curiosity ○ Perceived goal impediment ○ Redundant information ○ Perceived complexity ○ Lack of context ○ Unfamiliarity ○ Other: ----------	
Misapplying. ○ Misinterpretation ○ Mistrust ○ Confirmatory search ○ Rush understanding ○ Habits formation ○ Other: ----------	
Risk frequency and potential probability.	
Risk assessment. Please determines how acceptable is the risk on the trust calibration in the Human-AI task.	
No risks have been identified	
The risks require changing the XAI method	
The risks can be mitigated on the design level	
The risks have little effect on trust calibration	
Is there any additional information that you would like to be considered or you think is relevant?	

Fig. 6. Trust calibration risk assessment sheet – usability dimension.

conducted by Naiseh et al. (Naiseh et al., 2021) Table 3 shows C-XAI design principles, associated trust calibration risks and suggested design techniques. C-XAI also provides further information and examples of XAI interface design for related design principles in A6D2.

2.4. The evaluation phase

The final stage of the design process evaluates the effectiveness of the prototype(s) developed during the Selection and Implementation stage. The evaluation stage shall be an iterative process that involves a cyclical process of refining, testing the XAI design until it meets the needs and expectations of its users. It shall also meet the requirements for meeting threshold trust calibration risks. To assist this process, C-XAI proposes four behavioural metrics proposed in the literature to measure trust calibration risks during human-AI task e.g., (Buçinca, Malaya & Gajos, 2021; Chromik, Eiband, Buchner, Krüger & Butz, 2021), summarised in Table 4. Then, based on the feedback and the behavioural metrics, the design is refined, and the process is repeated until an optimal solution is achieved. Each iteration builds upon the insights gained from the previous cycle, allowing designers to continuously improve the XAI design. Once an acceptable solution is achieved, the design is ready to undergo more traditional evaluations using human factors (Whitefield, Wilson & Dowell, 1991) and performance analysis methods (Vermeeren et al., 2010).

We finally summarise the C-XAI process in Fig. 7.

3. Evaluating c-xai

C-XAI is a design method that offers a systematic approach to guide the design of XAI interfaces for trust calibration. The primary objective of C-XAI is to enhance the likelihood of building an XAI interface that effectively reduces trust calibration errors. Nevertheless, designing XAI interfaces for trust calibration presents challenges due to limited knowledge in the field and the dynamic nature of trust calibration. Additionally, understanding users' personalities and intentions for usage is crucial for effective design, further adding to the complexity. Therefore, the C-XAI framework does not claim to completely eliminate potential trust calibration risks but instead follows a systematic approach to assist the design team in anticipating and potentially mitigating such risks at the design stage. It considers both human and technical factors associated with the underlying XAI methods. This section builds upon the previous section, where C-XAI was introduced, and discusses how we gathered evidence on the potential of C-XAI to support the design process of XAI user interfaces with trust calibration in mind. We delve into the design of the C-XAI evaluation process, present the resulting findings, and draw conclusions based on them.

3.1. Evaluation goals

The objective of this evaluation is to assess the usefulness and effectiveness of the C-XAI framework in assisting the design team in identifying and addressing design requirements to mitigate potential trust calibration risks. In this evaluation study, we will use the effectiveness of a design process (as suggested by Veryzer and Borja (Veryzer & Borja de Mozota, 2005)) as a baseline for comparison. The effectiveness of a design process has four main dimensions:

- Aid the focus of the stakeholders on the design problem.
- Increase the awareness designing XAI interfaces with trust calibration goal and its diverse contexts and needs.
- Better design of the product
- Effective communication tool and increased engagement amongst the design team.

We also evaluate C-XAI documents based on the following criteria:

Table 3
C-XAI design principles, associated trust calibration risks and design techniques.

A6D1 Design principle	TITLE: Trust calibration design principles and guidelines Description	Recommended for risks	Potential design techniques
Persuasive design	XAI interface designers to increase users' tendency to engage with the explanation.	Skipping	- Reward. XAI interface that demonstrates the benefits and the rewards of engaging with the explanation has great persuasive learning powers. - Suggestions. XAI design that offers suggestions for material to read about the AI and its explanation will have greater persuasive learning powers.
Challenging habitual actions	XAI interfaces design for a calibrated trust may need to consider challenging users from developing habits with the XAI interface.	Skipping and misapplying	- Feedback on XAI knowledge. Lack of feedback might lead users to form habits and overestimate their understanding of the system. Designer may need to address this issue and necessitate tools to refresh users' actual knowledge and engage them in the learning process. - Friction design. An XAI interface design that include interaction that disrupt habitual and mindless automatic interaction, promote moments of reflection and more cognitive interaction.
Attention guidance	XAI interface to promote a desired user behaviour to look for relevant content in the explanation and combat overlooking it.	Skipping and misapplying	- Navigation. An XAI interface design that makes the explanation easy to find and navigate provides opportunities to persuade users to read the explanation. - Tunnelling cues. An XAI interface design that guides users through the process of reading the explanation. - Abstraction. An XAI interface design that fragments a complex explanation and presents it at multiple abstraction levels is more likely to convince users to read the explanation.
Training and learning	XAI interface design may need to train the users and facilitate their	Misapplying	- *Onboarding technique. An XAI interface design that provide an*

(*continued on next page*)

Table 3 (*continued*)

A6D1 Design principle	TITLE: Trust calibration design principles and guidelines Description	Recommended for risks	Potential design techniques
	understanding of the explanation method capabilities, limitations and learn optimal usage scenarios before and during the interaction with the XAI interface.		*onboarding exercise with the AI explanations and how to interpret them, this can include video tutorial or even a short course.* - *FAQ. An XAI interface design that includes Frequently Asked Questions on how to interpret an explanation output.* - Interactive. Interactivity. Interactive XAI design features that allow users to simulate potential outputs, compare, manipulate, and assess their impacts has been shown to support learning. - Self-learning tools. An XAI interface that provide tools to facilitate the process of learning from explanations, e.g., taking notes or archiving explanations for future interactions.

Table 4
Calibrated trust evaluation.

Calibrated trust evaluation – behavioural measures	Description	Reported studies
Agreement percentage	It refers to the percentage of trials in which the participants decided to agree with the AI-based recommendations.	(Naiseh et al., 2023; Yin, Vaughan & Wallach, 2019; Zhang et al., 2020)
Compliance percentage	It refers to the percentage of trials in which the participants choose to follow the AI-based recommendation. The main difference between Agreement and Compliance measures is that the participants agreed with the AI-based recommendation and automatically made the final decision in the agreement case. In contrast, compliance only considers the case where the participants disagree with the AI-based recommendation, but they intend to comply with the AI-based recommendation.	(Naiseh et al., 2023; Yin et al., 2019; Zhang et al., 2020)
Incorrect decisions	This is a team-performance measure, and it is extracted from incorrect decisions made between the human and the AI.	(Buçinca et al., 2021; Bussone et al., 2015; Lai & Tan, 2019; Yang, Huang, Scholtz & Arendt, 2020)
Correct decisions	It measures the percentage where the collaborative decision-making between the human and the AI has led to a correct decision.	(Buçinca et al., 2021; Lai & Tan, 2019; Naiseh et al., 2023; Yang et al., 2020)

- *Completeness.* This criterion refers to C-XAI ability to cover all design stages to develop the XAI interface. It also considers whether the guidelines provided to aid the design process are enough.
- *Understandability.* This criterion is to determine the understandability degree of C-XAI from the stakeholders' point of view. It also covers the evaluation of supporting documents and templates.
- *Usefulness.* This criterion aims to evaluate how the method and its supporting documents simplify and improve the XAI interface design process.

3.2. Design of the evaluation study

The evaluation study consists of two main phases. Evaluation material used during these phases can be found at [https://bit.ly/3gmdPpy].

3.2.1. Phase 1 – expert evaluation

The focus of this phase is to validate the C-XAI framework and its accompanying documents through the perspective of domain experts. The primary objective of this phase was to leverage the expertise of domain specialists to enhance the quality and effectiveness of the templates and supporting documents. Their valuable insights and recommendations served as a foundation for refining and improving these materials. The outcome of this phase plays a crucial role in ensuring that the templates and supporting documents are robust and aligned with expert perspectives, thereby facilitating the subsequent evaluation study.

3.2.2. Phase 2 - Case study evaluation

In this phase, we employed a case study approach to assess the effectiveness of C-XAI. Participants were tasked with designing XAI interfaces using the C-XAI activities and documents. We selected the Screening Prescription (SP) case study, which is an AI-based tool used to determine if a prescription is suitable for a specific patient based on their profile and medical history. This tool is designed to aid healthcare professionals in their decision-making process. We chose this case study because trust calibration is crucial for the successful implementation of the SP tool in real-world scenarios. Medical practitioners interacting with the SP tool may make trust calibration errors due to the dynamic nature of the AI's margin of error. The developers of the SP interface aim to create a safe and effective interface by ensuring transparency and explainability of the underlying logic. The C-XAI framework plays a crucial role in identifying the necessary explanations for the SP tool and identifying potential trust calibration risks. It also provides design principles at the XAI interface level to enhance the trust calibration process. The case study approach allows for an in-depth exploration of the phenomenon using multiple data sources. Our objective in the case study evaluation is to examine whether the proposed method assists various stakeholders in successfully developing innovative XAI interfaces that enhance trust calibration. By observing stakeholders' interactions with the C-XAI framework, we can gather rich data and gain a deeper understanding of their reactions during the design process. The case study will also help identify any challenges or shortcomings in the supporting documents and templates. The findings from the evaluation study will be used to refine and optimize C-XAI framework. The choice of the case study evaluation method is based on its inherent advantages, available resources, time constraints, and the nature of the research.

3.3. Phase 1: expert evaluation

This phase entailed the evaluation of C-XAI framework, along with its accompanying documents and templates, from the perspective of domain experts. The researcher first presented the templates and documents to experts who possessed relevant expertise in areas such as artificial intelligence (AI), human-computer interaction (HCI), requirement engineering, and psychology. Prior to commencing the evaluation,

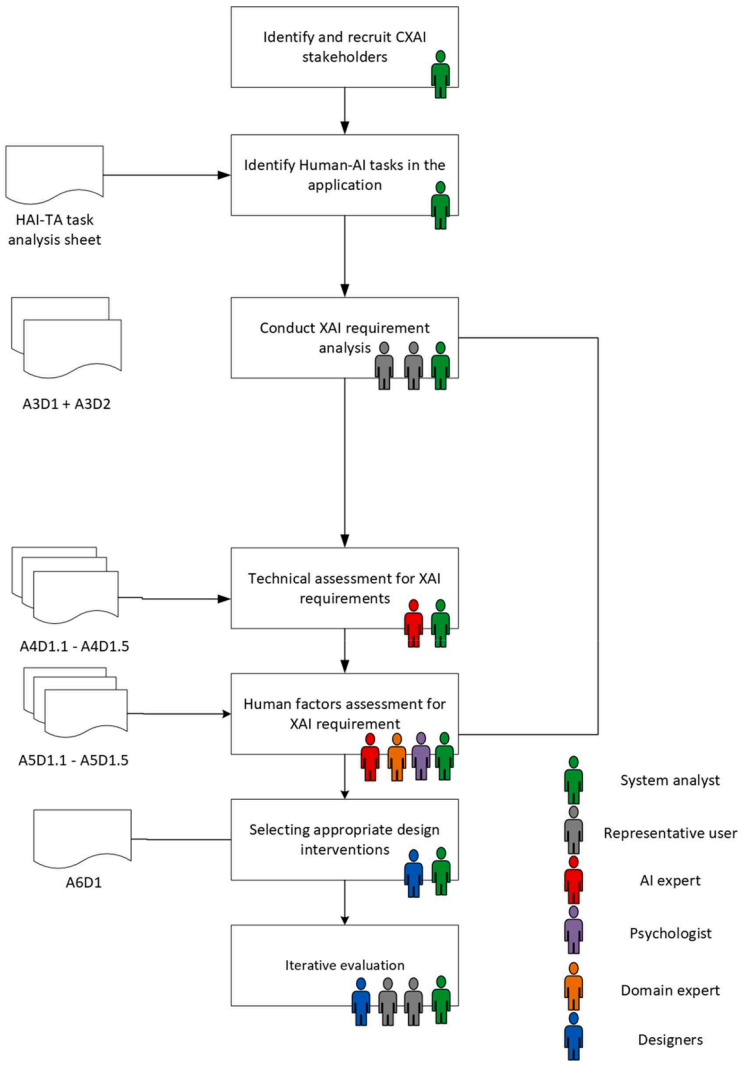

Fig. 7. C-XAI process summary.

a comprehensive 20-minute presentation was delivered by the researcher, offering a detailed justification for all the materials provided. The evaluation process involved a critical review and validation of the documents by the experts. They were tasked with carefully examining each document, scrutinizing its content and structure, and assessing its suitability and relevance. The experts were encouraged to provide constructive feedback and suggestions for refinement, allowing them the freedom to add or remove elements as they deemed necessary.

The selection of participants for this phase adhered to two specific inclusion criteria. Firstly, individuals were required to possess a minimum of five years of experience in relevant fields, including artificial intelligence (AI), human-computer interaction (HCI), requirement engineering, and psychology. This criterion ensured that participants had substantial expertise in areas closely related to the evaluation objectives. Secondly, participants were expected to have familiarity with the literature on trust or trust calibration, further reinforcing the credibility and validity of the findings.

In accordance with these inclusion criteria, six participants accepted to participate in the expert evaluation study. We outline our experts experience in the filed in Table 5. Prior to their involvement, experts were provided with a comprehensive booklet that contained detailed information about C-XAI framework. The booklet served as a reference guide throughout the evaluation process, offering participants insights into the methodology and facilitating a more informed evaluation.

During the evaluation, participants were provided with guided questions specifically designed to assess the effectiveness and suitability of C-XAI framework. They were encouraged to critically evaluate the method, identify strengths and weaknesses, and propose any necessary amendments or additions. Participants were asked to provide their feedback and comments directly within the booklet, enabling them to actively contribute to the refinement and enhancement of C-XAI framework.

By engaging participants in a structured evaluation process and soliciting their valuable input, this phase aimed to gather diverse perspectives and ensure the robustness and relevance of C-XAI framework. Expert feedback was collected in this phase and produced C-XAIv2; the second version of C-XAI. After collecting experts' feedback, the researcher further refined the templates and the supporting documents.

Table 5
Demographics for Experts who participated in the expert evaluation study.

Participants ID	Background and expertise	Experience
P1	Machine learning expert	8 years
P2	Requirement engineering expert	5 years
P3	Interaction design and UX expert	6 years
P4	Behavioural change/ Psychology	10 years
P5	Software engineering	8 years
P6	Human-computer interaction expert	12 ears

3.4. Phase 2: case study evaluation

In Phase 2, the objective was to utilise C-XAI framework and its accompanying supporting documents to design an XAI interface for SP tool. The process commenced with participants being invited to engage in focus group discussion, where they had the opportunity to deeply understand the design problem and exchange their viewpoints regarding how the XAI interface design could effectively facilitate trust calibration. Following the initial focus group session, participants were randomly assigned to two separate groups. Each group was tasked with designing an XAI interface for SP tool, employing C-XAI framework as a guiding framework. To aid them in this process, participants were provided with three essential documents:

- **HAI-TA document for SP tool:** The completed task analysis sheet for capturing XAI requirement. It served as a comprehensive guide, outlining various XAI requirements for the SP tool. This document was based on earlier elicitation by IQemo IQHealthTech[1] for their SP tool. IQemo is a partner and funded the project. That means the Phase 2 evaluation study evaluated C-XAI framework started from the Assessment phase, i.e., the identification process has already been performed by the IQHealthTech organisation. IQHealthTech has conducted its XAI requirements having their system analyst and medical practitioners from two hospitals in the UK as representative users. The available resources and time informed the choice of such an approach.
- **Guidance document.** It provides an overview of the design goal, design problem, and the specific Human-AI decision-making task under consideration (in this instance, the SP tool).
- **C-XAI documents,** which outline the systematic approach and guidelines for designing explainable artificial intelligence (XAI) interfaces. These documents offer comprehensive information on various aspects, such as eliciting explainability needs, assessing XAI requirements, and best practices for trust calibration.

Two design sessions were conducted to design an XAI interface for SP tool. The researcher started both sessions by briefing the session's aim and a description of the design problem. The system analysts who guided the design process were familiar with requirement engineering standards and guided different stakeholders through the design process. Various stakeholders worked in two groups to evoke brainstorming and critical thinking. Participants worked together to produce the prototypes following the instructions provided in study material.

After completing each session, participants were given an open-ended survey to gather participants experience during the design sessions. They were also asked whether they had encountered issues or difficulties during the design process. This survey aimed to take participants overall experience into consideration when conducting the analysis. The collected data was analysed to refine the method templates and supporting documents. The researcher focused on evaluating the templates' effectiveness, i.e., to what extent the method and its template helped participants understand and analyse the design problem. The researcher also focused on identifying templates' flaws and disadvantages.

3.4.1. Participants

Participants involved in this stage of evaluation were not engaged in any earlier studies that helped develop C-XAI framework. The main aim of this inclusion criterion was to ensure that participants had no prior knowledge about C-XAI framework . Participants who agreed to participate in this stage collaborated with the system analyst to generate the XAI interface design. We chose our AI expert to have experience with both AI and XAI knowledge. Our argument is that AI experts with mix expertise of AI and XAI can assess not only the interpretability features of the XAI algorithm but also its overall performance, efficiency, and integration with other AI components. Table 6 present the demographic information about the participants involved in the case study evaluation. Participants selection criteria for the evaluation study included (i) representative users who have experience in prescription classification, and (ii) experts shall have at least 5 years of experience in their domains. The evaluation study employed a convenience sampling approach.

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

3.5. Findings

This section presents the findings from the evaluation study. The results will be discussed based on the evaluation study goals: completeness, understandability, usefulness and effectiveness and divided in two sections – expert evaluation and case study evaluation.

3.5.1. Expert evaluation findings

As a general observation, experts rated the proposed method as a practical, comprehensive and complete method for designing XAI interfaces. Experts mentioned that the templates provide a comprehensive description of the design process and help C-XAI users identify several trust calibrations risks. However, five out of six experts mentioned that examples and heuristics to help C-XAI user to map between XAI technical properties and trust calibration risks should be provided as an additional document to C-XAI framework . They commented that such a document would help stimulate the thinking process and evoke brainstorming around the design problem. It is also to facilitate a dialogue between different stakeholders during the design process.

Considering the content of the templates, AI experts suggested changing the answers of some XAI method properties that have "*Yes/No*" options to "*low/medium/high*" options. This would represent an accurate description of the property of the XAI method. P1, a machine learning expert, mentioned: "*all explanation models have these features, but it depends on their levels … so yes/no answers are not really descriptive*". Five XAI technical properties were changed accordingly: *Complexity, Novelty, Coherence, Soundness and Completeness*.

Table 6
Stakeholders' evaluation.

Participants ID	Background and expertise	Experience	Role	Region
P1	Requirement engineer	9 years	System analyst / Industry	UK
P2	Medical doctor	6 years	Representative users/ Industry	UK
P3	Medical doctor	6 years	Representative users / Industry	UK
P4	Expert doctor	10 years	Domain expert / Industry	UK
P5	UX designer	5 years	Designer / Industry	UK
P6	Psychologist	7 years	Psychologist / Academia	UK
P7	Machine learning engineer	9 years	AI expert / Industry	UK
P8	Machine learning researcher	12 years	AI expert / Academia	Germany
P9	Pharmacists	6 years	Representative users / Industry	UK
P10	Pharmacists	8 years	Representative users / Industry	UK
P11	Expert pharmacist	11 years	Domain expert / Industry	UK
P12	HCI researcher	7 years	Designer / Industry	US
P13	Requirement engineering researcher	6 years	System analyst / Industry	UK
P14	Psychology researcher	9 years	Psychologist / Academia	Italy

Regarding the terminology used in the templates and the method. Participants agreed that the language and the terminology are generally understandable. However, some participants have suggested adding extra document to define the technical properties of XAI methos and trust calibration risks provided in the templates. In addition, they mentioned that some of the provided terminologies are not self-explanatory and might cause a misunderstanding to the C-XAI stakeholders. For instance, P4 mentioned, "*Well … misusing the explanation could be interpreted differently… it is better to provide what does that mean*". Therefore, a glossary template was added to the methods' supporting documents.

Regarding the technical assessment sheets, participants suggested that some elements can be removed because they are repetitive. The researcher argued that this was adopted from a framework in the literature, but the experts made the case that providing such repetitive questions could confuse the stakeholders. For instance, actionability of XAI method appeared both in the operational assessment sheet and usability assessment sheet. Furthermore, experts suggested adding five different trust calibration risk assessment sheets corresponding to each XAI assessment dimension. Participants argued that it might be overfolding and confusing to combine five assessment dimensions in one template. Five Trust calibration risk assessment sheets were developed accordingly.

Experts were interested in the good design practice document and agreed that such information is useful during the XAI interface development, "*I like the guidelines … designers and requirement engineers would definitely need them*". However, amendments were suggested related to the terminology, length and writing style of the guidelines. Three experts mentioned the length of the guidelines and suggested shortening them. P3 suggested, "*The guidelines need to be shortened as they contain a lot of information, and it might be difficult for the designers and system engineers to follow with it*". Two experts also described the guidelines as academic guidelines where the guidelines of C-XAI framework shall be more informative to the designers. P4 added, "*Well I can understand these guidelines, but I doubt designers would be able to follow up with this style of writing… it is more academic*". Therefore, the guidelines and good design practices were amended and styled to reflect their feedback. Table 7 presents main positives and negatives received from our expert's evaluation in terms of Completeness, Understandability and Usefulness of C-XAI. As a result of this stage a second versions of C-XAIv2 have been developed.

Table 7
Expert's evaluation of completeness, understandability and usefulness of C-XAI.

Completeness	
Templates and supported documents covered all the required assessments related to enhancing trust calibration problem "*I would say it is complete, I cannot think of anything missing*". Another remarked that "*The entire process was explained sufficiently in the booklet*".	+
Participants required additional spaces to improve the communication between the stages.	–
Participants asked for additional supporting documentation that explains the main terminologies used in the design method, Participant stated, "*what does explanation completeness exactly mean*".	–
The evaluation stage missed a direction to help C-XAI users measure cognitive thinking and correct interpretations. These measures were added to C-XAI iterative evaluation.	–
AI experts suggested changing the answers of some XAI method properties that have "*Yes/No*" options to "*low/medium/high*" options.	–
Experts suggested adding five different trust calibration risk assessment sheets corresponding to each XAI assessment dimension.	–
Understandability	
The templates and the workflow provided a clear structure of the C-XAI framework, which helped the system analyst.	+
The order of the stages was logical, for instance, providing an assessment for the explanation method before the trust calibration risk assessment.	+
Participants agreed that more examples are required in each step to enhance users' understandability of the design problem, e.g., P3 during the trust calibration risk assessment stated, "*I would like to see more examples for explanation properties that may introduce trust calibration risk*".	–
System analysts mentioned that good design practice for eliciting explainability needs pointed out several requirements to be considered. For instance, the system analyst was not aware of collecting explainability needs before and after consuming the main explanation.	+
Explaining 'when' and 'how' to implement the proposed design guidelines stimulated participants thinking and helped them to come up with innovative design solutions.	+
Usefulness	
Experts mentioned the templates encapsulate different information in relation to trust calibration design problem.	+
Experts mentioned that the generated design would help reduce the cost of users' errors, specifically when a high-stake Human-AI task is implemented. They mentioned that the C-XAI framework will help the designers recognise what those potential problems are to be solved in the design to increase the effectiveness of the explanation. One expert mentioned, "*I can see how this method would point out different user errors … methods such as User-centred design may not be able to reveal such problems … it takes a considerable amount of research to come up with these ideas and errors*".	+
Experts were interested in the design guidelines and good design practice and agreed that such information is useful during the XAI interface development	+

3.6. Case study evaluation

Participants utilised C-XAI in the design process to design an XAI interface for SP tool. Participants completed a questionnaire consisting of two sections, the first section related to evaluating the effectiveness of C-XAI during the design process. While the second section is related to the potential improvements that can be performed to the method. In general, all participants emphasised that the C-XAI was useful and effective in the design process. In this section, we report results from our analysis and discuss the benefits and risks of using the method that emerged from the questionnaire analysis. As a results of this evaluation, version C-XAIv3 has been developed.

3.6.1. Benefits of using C-XAI in the design process

The following points discuss the benefits and final amendments made to C-XAI framework based on researcher observations and questionnaire analysis.

3.6.1.1. Effective communication and increased engagement. C-XAI was evaluated in terms of facilitating effective and clear communication between different stakeholders. Adopting participatory design and specifying stakeholders' roles and tasks during the design process provided a clear direction to each participant. Due to the fact the current methods in the literature such as Eiband et al. (2019) lacked a straightforward method to address both technical and human factor aspects related to XAI methods, C-XAI led to an increase in the engagement of different stakeholders during the design process. However, some communication issues were revealed during the evaluation based on the researcher observations and participants' feedback to enhance the ease of using the method. For instance, participants required open-ended space in each template to add general notes and recommendations to the design method's subsequent activities. They argued that adding extra space in each template would enhance the communication between different design methods. Also, participants wanted to record observations and point them out to stakeholders in the next activity. For instance, the AI expert insisted on informing the design team about the importance of helping users interpret LIME explanations. Participants mentioned, "*Designer must know that these explanations are recommendation-specific explanations*".

Regarding the time needed to complete the technical assessment sheets, it was observed that AI experts required more time to provide complete and comprehensive feedback to each explanation method. In SP case study, one XAI method assessment activity was completed. This was not a challenging task to complete for the AI expert, but the system analyst had to provide some help by searching for relevant literature and

providing access to research databases. Consequently, it is recommended that the AI expert in this stage may need to complete this stage in multiple days or multiple sessions. This would ensure that the XAI technical properties are assessed precisely. P8 mentioned, "*assessing the explanation method need more time so the AI team in any software development would need multiple sessions across several days to collect information about the method and provide accurate assessment*".

3.6.1.2. Increased focus on the design problem. C-XAI framework is meant to increase the design team's focus on potential risks during the interaction between users and the XAI interface. It is meant to help the design team and different stakeholders in having a shared understanding of the potential users' errors. It is also to make the XAI interface effective in mitigating these errors. The observer indicated that the C-XAI framework helped participants understand the problem from actual context rather than their assumptions. It was noted that C-XAI templates and recommendations acted as a reference point to facilitate brainstorming activity on design XAI interfaces with trust calibration goal in mind. Most responses in the questionnaire expressed that C-XAI helped increase the focus on trust calibration problem and removed the focus from dealing with the explanation as informational content only, e.g., "*the method reflects the reality of trust calibration problem and have a link to many explainable AI literature and human behaviour as well … it helped me to focus my attention on combining these three areas together*".

3.6.1.3. Increased empathy toward end-users. C-XAI is meant to help the design and development team to build understanding towards the users and recognise their errors, needs and context. The term empathy means understanding and predicting people behaviours and psychological states. The observer indicates that using C-XAI templates allowed the design team and different stakeholders to identify and feel empathy for the end-users and understand their errors. With C-XAI framework in mind, creative design ideas were proposed with the aim of supporting users in calibrating their trust. Fig. 8 shows an example of the designs during the case study evaluation. Participants solved one trust calibration risk related to misinterpretation by including an additional view in the interface, which included the learning material that helps users of XAI interpret the explanations. The design show how participants adopted a learning design principle where they added an information icon next to each data feature in LIME explanation. The goal was to support users in interpreting the importance value of the data feature in the AI decision. Most responses in the questionnaire expressed that the C-XAI was useful and effective to react to and empathising with users' errors, e.g., the method templates enabled the design team to closely understand the user errors and behaviours that triggered the designers to set solutions. P1 stated, "*I found method shorten the distance between designers and real users*".

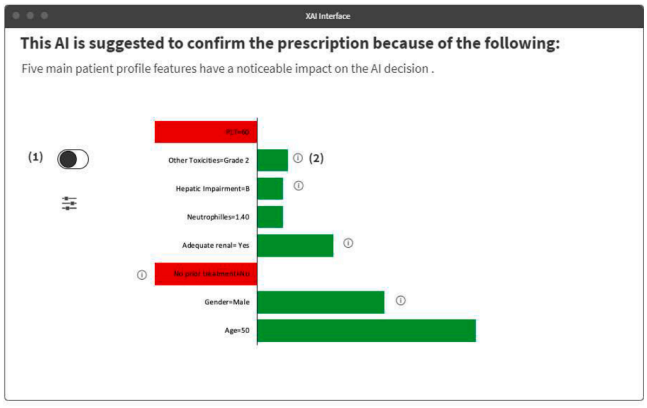

Fig. 8. Sample of participants designs during the design sessions.

3.6.1.4. Better software product design. C-XAI is meant to help produce better designs with a more valuable set of design features and make design decisions more informed. Understanding the potential behaviour and users' errors can facilitate new ideas and features, informing better design. During the evaluation, the observer indicates that the designers focused on specific trust calibration risks and tried to mitigate those risks. This helped the designers make better decisions by focusing on a specific problem and referring them to support their design ideas. Most responses in the questionnaire expressed that C-XAI was valuable and practical for better software design, e.g., "*using this method designers would set the right features for the right problem and risk*".

3.6.2. Drawbacks of using C-XAI framework in the design process

In this section we report several drawbacks identified by C-XAI users and researcher observation. We also elaborate how we addressed each of the risks in the version 3 of C-XAI.

- The method did not provide sufficient information about the nature of trust calibration risks which made participants struggle and rely on the psychology expert during the trust calibration risk assessment activity, P6 struggled to contextualise rush understanding risk and indicated, "*I was not really able to reflect how people would rush understand an explanation … I expected more elaboration*". A supporting document was added to the method to elaborate on each of the potential trust calibration risks.
- Five participants mentioned that the amount of effort to complete the templates could increase exponentially when the number of tasks and XAI method increase. This led them to recommend future recommendations to the method, e.g., ensure that an adequate number of stakeholders should be recruited based on these parameters. Therefore, C-XAI v3 included a recommendation to the system analyst to consider this point during the participants' recruitment process.
- Trust calibration risks include users' errors and potential cognitive biases during Human-AI collaborative decision-making tasks. Also, the method included templates that reveal several limitations of the XAI method. This information facilitated the stakeholder to make assumptions about the end-users and XAI methods and they may generalise these assumptions over other XAI methods or tasks.
- C-XAI booklet was confusing to participants. Many participants had to ask questions to the system analyst to guide them through the booklet. Many responses in GAQ asked to split the documents into several documents, where each document is only relevant for the current activity. Other participants were also asked to specify each stage's input and output at the beginning of each activity. P5, a UX expert, mentioned, "*the booklet contains so much relevant information for each activity … I would recommend splitting it into multiple files*".
- C-XAI is not intended to be used in all scenarios but is suitable for medium- and high-stakes decisions where human decision-makers are ultimately accountable, and low-frequency decision-making where the decision-maker has time to engage with AI explanations.

4. Conclusion

In this paper, we introduce C-XAI (Calibrated trust for eXplainable AI), a design method that aims to facilitate collaboration between various stakeholders from different disciplines in order to create XAI interfaces with a focus on trust calibration. The method comprises four key stages. The first stage, identification, involves identifying representative users and stakeholders who can provide insights into the requirements for explanations relevant to human-AI tasks. In the assessment stage, these explanation requirements are evaluated from both technical and human perspectives to identify potential trust calibration errors. The assessment results offer valuable insights to the design team and system analysts, aiding their understanding of the design problem and enabling them to propose suitable solutions in the

Selection and Implementation phase. Throughout this phase, a multi-disciplinary team of experts engages in activities aimed at mitigating potential risks associated with the C-XAI process. The iterative evaluation stage concludes with follow-up meetings to reassess the design and analyse potential trust calibration risks. C-XAI underwent two iterations of evaluation to develop the final version, C-XAIv3. In the first iteration, experts from various disciplines assessed the completeness, understandability, and usefulness of C-XAI and its supporting documents in the design process. Subsequently, different stakeholders and users were invited to utilize C-XAI to create XAI interfaces with trust calibration as a goal. It was observed that C-XAI facilitated effective communication between stakeholders and enhanced engagement in the design process. Furthermore, the results demonstrated that C-XAI aided users in recognizing trust calibration as a design challenge and fostering empathy towards the users of XAI interfaces.

Authors' contributions

MN: Conceptualisation, C-XAI development and writing of the initial draft

AS, BZ, DA: Critical review and Writing.

GR, NJ, RA: Critical review and Supervision

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EP/V00784X/1).

Statements related to data availability and ethical consideration

- This study obtained research data from publicly available sources.
- Ethical approval was obtained from Bournemouth University Ethics committee in accordance with relevant guidelines/regulations.

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.

Appendix A

Table A1
SD2: Guidelines during the elicitation of XAI requirements.

SD2: Guidelines during the elicitation of XAI requirements
Design for task-centred explanation rather than model explanation.
When explainability is employed in a collaborative decision-making environment, it is crucial to integrate it with the task workflow and task constraints. The explanation of the algorithm's logic should be closely intertwined with the subject at hand, specifically the task for which the recommendation and explanation are provided. In our user studies, we encountered an example of task constraint in Counterfactual explanation scenarios, where the explanation only offered information about the types of modifications that could be made to an AI recommendation to alter the decision. Our participants expressed concerns, noting that certain data features in the patient profile, such as patient demographics, had static values and could not be changed. In the process of eliciting users' mental models, the system analyst may need to place emphasis on eliciting the task constraints related to the selected explanation methods.
Integrate domain-related contextual information and assurances in the XAI interface.
Users who interact with technology desire the ability to validate its reliability before utilizing it. Similarly, users of XAI interfaces may request contextual information regarding the explanation and recommendation to assess its validity. For instance, during our user studies, some participants expected additional meta-information when explaining AI-based decisions in order to evaluate the tool's validity within legal and organizational boundaries. To enhance the reliability of explanations, appropriate assurances can be elicited during the elicitation process. For example, a validation stamp could be provided to the explanation, indicating that it has been reviewed by a domain expert during the development phase. Informing users about the validity of the explanation and its applicability to the current task can encourage them to utilize the explanation, thus improving their mental model.
Pre and Post XAI for calibrated trust.
To facilitate a correct interpretation of explanations, it is beneficial to develop a taxonomy, such as a help page, FAQs, chatbot, or training materials, that addresses users' questions regarding explanation usage scenarios, expert judgement on explanations, and the explanation development process. Our data also revealed that incorporating such features in the explainable interface can address unfamiliarity and the need for transparency. In terms of post-explainability, participants expressed a desire to validate the accuracy of the explanation by including a meta explanation or additional validation information. For example, those who questioned the correctness of the confidence score requested information about the sources of uncertainty and the margin of error associated with the confidence score. This approach is similar to the technique employed by Ads explainability in social media, which provides further explanations to users about the information inferred and used for targeting. However, it is important to consider privacy concerns that may arise due to the provision of additional explanations. This finding raises a research question regarding the information users may enquire about after being presented with the main explanation and in what specific situations. Addressing these questions can lead to the development of effective trust calibration interfaces by identifying the conditions that necessitate post-explainability. During the mental model elicitation process, the system analyst should be attentive to the various questions that participants may pose after utilizing the main explanation.
Enable effective tailoring.
Tailoring refers to the modification or explanatory process aimed at addressing user-specific questions, such as what-if scenarios. One notable risk associated with the absence of tailoring and personalization mechanisms is that the explanation may not always align with the needs of end-users, potentially leading to overlooking or misapplication. Tailoring can be employed in the XAI interface to account for the properties of the explainable method or address trust calibration risks. The literature review section in this thesis presents a comprehensive taxonomy for tailoring explanations to end-users, built upon a six-dimensional model of personalization.

Table A2
C-XAI provides a supporting document called SD2, which offers best practices for the system analyst during the elicitation of XAI requirements.

A3D1	SHEET TITLE: Methods to elicit XAI requirements
1	**Method:** Think-aloud problem solving, where participants think-aloud during a decision-making task. ***Strengths:*** It offers rich information about the users' mental model. ***Weaknesses:*** The process of transcription and data analysis might be time consuming.
2	**Method:** Think-Aloud protocol with specific question and answering activities. ***Strengths:*** It enables the researcher to target specific issues during the decision-making task. ***Weaknesses:*** It might introduce bias as it depends on the researcher skills in designing the study.
3	**Method:** Card Sorting based on the semantic similarity between the domain concepts ***Strengths:*** It enables the researcher to understand the relation between different domain concepts. ***Weaknesses:*** The collected data might be sparse about the events or processes.
4	**Method:** Nearest Neighbour task, where the participants select the best explanation that fits their task. ***Strengths:*** It can provide a quick understanding to the users' mental models. ***Weaknesses:*** It might be prone to the phenomena of Illusion of Explanatory Depth; people overestimate their understanding of complex systems.
5	**Method:** Self-explanation task, in which participants are presented with a number of AI-based recommendation and are asked to explain these recommendations. ***Strengths:*** It can provide quick access to users' mental model. ***Weaknesses:*** It requires a clear rationale for the choice of the AI-based recommendations to be the focus of the users' mental model elicitation task.
6	**Method:** Glitch Detector Task, in which participants are asked to identify the strengths and weaknesses in each of the available explanations. ***Strengths:*** It can support users to freely express their mental model that might be incorrect. ***Weaknesses:*** The glitches shall be built- in the design. Also, knowledge shields may reduce the awareness of the glitches.
7	**Method:** ShadowBox task, in which participants compare their understanding of the system to the expert explanation. ***Strengths:*** It can provide a quick access to users' mental model. ***Weaknesses:*** Participants may not be able to understand the expert explanation.

Heuristics of mapping between XAI method technical properties and trust calibration risks

Explanation property and value	Description	Potential trust calibration risks
Novelty=high	Explanation ability to generate surprising or abnormal information.	Habits formation Lack of curiosity
Complexity=high	Explanation level of detail and level of knowledge required to interpret the explanation.	Confirmatory search Perceived complexity
Completeness=Low	Explanation ability to generalize well beyond a particular recommendation.	Misinterpretation
Soundness=Medium	Explanation ability to be consistent and aligned with the underlying model.	Lack of context Mistrust
Interactivity = No	Explanation ability to be controllable and customizable to fit users' needs.	Lack of curiosity Lack of context
Explanation audience= AI expert	The audience that the explanation method was developed for is for AI experts.	Misinterpretation Perceived complexity

References

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 2429–2437).

Boehm, B. W., & DeMarco, T. (1997). Software risk management. *IEEE Software, 14*(3), 17.

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* (pp. 160–169). IEEE.

Cao, L. (2022). Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR), 55*(3), 1–38.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces* (pp. 307–317).

Clark, J. R., et al. (2022). Industry Led Use-Case Development for Human-Swarm Operations. *arXiv preprint arXiv:2207.09543*.

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

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence And Law, 28*, 415–439.

Harding, R. (1998). *Environmental decision making*. NSW, Australia: The Federation Press.

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

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).

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).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence, 267*, 1–38.

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

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: when design meets trust calibration. *World wide web, 24*(5), 1857–1884.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal Human Computer Studies, 169*, Article 102941.

Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendations and calibrated trust: two systematic user errors. *Computer, 54*(10), 28–37.

Naiseh, M., Jiang, N., Ma, J., & Ali, R. (2020). Explainable recommendations in intelligent systems: delivery methods, modalities and risks. In *Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings 14* (pp. 212–228). 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).

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

Sokol, K., & Flach, P. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 56–67).

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review, 28*(2), 454–475.

Vermeeren, A. P. O. S., Law, E. L.-C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries* (pp. 521–530).

Veryzer, R. W., & Borja de Mozota, B. (2005). The impact of user-oriented design on new product development: An examination of fundamental relationships. *Journal of Product Innovation Management, 22*(2), 128–143.

Wagner, A. R., & Robinette, P. (2021). An explanation is not an excuse: Trust calibration in an age of transparent robots. *Trust in Human-Robot Interaction* (pp. 197–208). Elsevier.

Whitefield, A., Wilson, F., & Dowell, J. (1991). A framework for human factors evaluation. *Behaviour & Information Technology, 10*(1), 65–79.

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., Vaughan, J. Wortman, & 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.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731.
Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305).