Investigating Human-Centered Perspectives in Explainable Artificial Intelligence

Muhammad Suffian[1,*], Ilia Stepin[2], Jose Maria Alonso-Moral[2] and Alessandro Bogliolo[1]

[1]Department of Pure and Applied Sciences, University of Urbino Carlo Bo, Urbino, Italy
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain

Abstract

The widespread use of Artificial Intelligence (AI) in various domains has led to a growing demand for algorithmic understanding, transparency, and trustworthiness. The field of eXplainable AI (XAI) aims to develop techniques that can inspect and explain AI systems' behaviour in a way that is understandable to humans. However, the effectiveness of explanations depends on how users perceive them, and their acceptability is connected with the level of understanding and compatibility with users' existing knowledge. So far, researchers in XAI have primarily focused on technical aspects of explanations, mostly without considering users' needs, and this aspect is necessary to consider for a trustworthy AI. In the meantime, there is a growing interest in human-centered approaches that focus on the intersection between AI and human-computer interaction, what is termed as human-centered XAI (HC-XAI). HC-XAI explores methods to achieve user satisfaction, trust, and acceptance for XAI systems. This paper presents a systematic survey on HC-XAI, reviewing 75 papers from various digital libraries. The contributions of this paper include: (1) identifying common human-centered approaches, (2) providing readers with insights into design perspectives of HC-XAI approaches, and (3) categorising with quantitative and qualitative analysis of all the papers under study. The findings stimulate discussions and shed light on ongoing and upcoming research in HC-XAI.

Keywords

Artificial Intelligence, Explainable AI, Human-centered XAI, XAI Design Perspectives, Systematic Survey

1. Introduction

Over the last two decades, Artificial Intelligence (AI) has received an overwhelming response from daily life applications, and industries such as autonomous vehicles, financial services, and healthcare [1, 2]. As the utility of AI systems in every walk of life has increased, the call for algorithmic understanding, transparency, and trustworthiness has become a matter of regulation [3]. The European Union General Data Protection Regulation (GDPR) [4] refers to the "right to explanation" of European citizens when their personal data are automatically processed by AI systems. Moreover, the amendments to the new AI Act adopted at the first reading

*XAI.it 2023 - Italian Workshop on Explainable Artificial Intelligence, November 08, 2023 | co-located with AI*IA 2023*
*Corresponding author.
m.suffian@campus.uniurb.it (M. Suffian); ilia.stepin@usc.es (I. Stepin); josemaria.alonso.maral@usc.es (J. M. Alonso-Moral); alessandro.bogliolo@uniurb.it (A. Bogliolo)
0000-0002-1946-285X (M. Suffian); 0000-0002-4508-7555 (I. Stepin); 0000-0003-3673-421X (J. M. Alonso-Moral); 0000-0001-6666-3315 (A. Bogliolo)

in June 2023 [5] include the newly proposed Art. 68c "A right to explanation of individual decision-making" which explicitly requires AI systems to be endowed with different levels of transparency depending on the risk of their application; being mandatory to provide explanations for algorithmic decisions when requested in case of high-risk applications. Accordingly, transparency is considered to be a prerequisite for explainability. Thus, in eXplainable Artificial Intelligence (XAI), researchers aim to develop techniques to inspect and explain the working of AI systems that humans can understand easily [1]. The goal is to increase the transparency, understandability, and usability of such intelligent systems.

Notice that, the development of XAI applications is complex because the effectiveness of an explanation depends on how the user perceives it [6]. Making the model transparent does not guarantee it will be fully understandable at the user end [7]. Hence, understandable explanations are necessary, and the quality of an explanation is determined by its ability to provide relevant information that can be understood and applied. Also, it should be compatible with the user's knowledge and needs [8]. However, researchers in the field of XAI have primarily focused on the technical perspective of explanations, giving little attention to users' needs [9, 10]. Many algorithms and techniques are designed based on researchers' intuition of what makes a "good" explanation, but without considering its usability, effectiveness, and practicality for human users [11].

In recent years, the XAI community has shown a growing interest in human-centered approaches [12, 13, 14, 15, 16]. To address the crucial requirements regarding transparency, understanding and trustworthiness in XAI systems, a new area of research has emerged where AI and human-computer interaction (HCI) intersect [17]. This research field is termed human-centered XAI (HC-XAI) [18]. It is important to examine the existing theoretical and practical frameworks for generation and evaluation of explanations, regarding both effectiveness and impact of explanations on user's trust and satisfaction. The consideration for user requirements and contextual factors in the design and evaluation of explanations are deemed necessary [1], while tailoring explanations to specific user goals and preferences is challenging [19]. Numerous review articles on XAI have been published, exploring different aspects and highlighting the challenges and potential research directions. Some surveys briefly touch upon evaluation measures within a broader perspective of XAI [20, 1] or mainly discuss evaluation through user studies [21, 22, 23], others have a narrower scope focusing on specific application domains or subareas of XAI [17, 24, 18]. However, despite these efforts, the attention given specifically to HC-XAI was limited and scattered, lacking a dedicated focus on the subject.

In this paper, we conducted a systematic survey on HC-XAI. In contrast to previous surveys, our approach paid major attention to research papers dealing with human-centered factors, including artefacts and techniques which strive to address the human needs for explanations. The survey focuses on reviewing 75 papers from Scopus, Web of Science, IEEE, and ACM digital libraries. Our objectives are: 1) to lay out common approaches for HC-XAI, 2) to provide insights on the design perspectives of common approaches, and 3) the categorisation of included papers. To achieve these objectives, we conducted a systematic literature review driven by the following specific research questions:

- **RQ1:** What are the most common human-centered approaches (including frameworks, methods, theories, and artefacts) in XAI?

- **RQ2:** How is the co-design perspective considered in HC-XAI approaches?
- **RQ3:** How grounded are human-centered approaches on theoretical and practical aspects of explanations?

Accordingly, we first categorise and provide insights into common approaches for HC-XAI. Then, we do quantitative and qualitative analyses of the papers under study. Our overarching goal is to bring attention to the ongoing and upcoming research related to HC-XAI. Therefore, the rest of the paper is organised as follows. Section 2 introduces background and related work. Section 3 details the methods used in the systematic literature review. Section 4 presents the most outstanding quantitative results. Section 5 provides a qualitative analysis of the most relevant papers. Section 6 discusses implications and research opportunities. Section 7 provides readers with final remarks.

2. Background and Related Work

Human-centered explanations are designed and tailored to meet the needs, understanding, and cognitive abilities of human recipients [23, 18]. These explanations aim to enhance the comprehensibility and accessibility of complex information for individuals by considering their unique perspectives, backgrounds, and experiences. The central concept driving human-centered explanations is to bridge the gap between specialised (technical) knowledge and the broader (or specific) target audience. By prioritising the human recipient, explanations are easier to understand, facilitate informed decision-making, and foster transparency, trust, and engagement between experts and the wider public [25, 17].

On the one hand, there are numerous reviews available on the expanding field of XAI. They play a vital role in defining and establishing the concept of XAI [26, 1]. They delve into exploring the interconnections between XAI and related fields of study [26, 15], categorising different methodologies [24, 27], analysing the user's perspective [21], examining evaluation practices [28], and proposing future directions for research [26, 1]. However, it is worth noting that the importance of human-centered approaches in the context of XAI is only briefly acknowledged in previous reviews.

On the other hand, in recent years, there is a growing interest to highlight the needs and importance of human-centered approaches [12, 13, 14, 15, 16]. To address this aspect, Chromik and Schuessler [29] proposed a taxonomy that incorporates human perspectives for evaluating XAI. Additionally, Lai et al. [25] conducted a comprehensive review of studies focusing on collaborative human-AI decision-making. Their review specifically explores the role of explanations in evaluating the success of such collaborative efforts. Furthermore, Ferreira and Monteiro [21] paid attention to the user experience in XAI applications. They carried out an analysis of the users themselves, their motivations, and the contextual factors that influence the presentation of explanations.

In the following sections we will provide readers with a systematic literature survey to categorise existing approaches which introduce, devise, or apply human-centered artefacts and techniques for XAI.

3. Methods

In this section, we outline our methodology for identification, screening, eligibility (reviewing), and inclusion of papers. To ensure comprehensive insights into the HC-XAI domain, we conducted a systematic collection of papers, regarding both quantitative and qualitative information. Our methodology follows the guidelines established by Moher et al. [30] for preferred reporting items for systematic reviews and meta-analyses (PRISMA).

Publication Search Engine. In order to ensure a comprehensive and manageable selection of papers, we selected interdisciplinary databases (search engines) such as Web of Science (WoS), Scopus, ACM Digital Library, and IEEE digital Explorer for the initial selection of papers. These databases not only encompass research publications in the field of Computer Science, but also index studies in scientific fields such as AI, HCI, Social Sciences and Philosophy; enabling a comprehensive review of the literature related to our research questions.

Publication Year. We limited our selection to papers published between 2018 and mid-May, 2023. This time frame was chosen due to the significant surge of interest in XAI in recent years, particularly following the implementation of regulations such as the GDPR and AI act in Europe, as well as other AI-related acts in the UK and the US.

Search Strategy. On May 15th, 2023, we conducted the advanced search using web tools provided by the selected digital databases. These tools allow for the replication of the study and ensure consistent querying across all the databases. We performed a query consisting of multiple terms on the title, abstract, and author keywords.

The query for the digital databases was as follows: ((expla* AI OR expla* Artificial Intelligence OR Counterfactual expla*) OR (interpret* AI OR interpret* Artificial Intelligence)) AND (generat* OR framework* OR develop* OR (human-centered OR human-centred) OR (user-centered OR user-centred)). To maximise the diversity of the retrieved papers, we used word-stems in our search query. For example, the search item "expla*" covered all word-forms such as "explanation", "explaining", "explanatory", etc. We used the search terms to gather the most recent publications that mention explainable AI and interpretable AI, respectively, across all subject areas. Also, the search query was overlapping to differentiate between publications covering human-centered and user-centered development frameworks (to cover the broader perspective of human-centricity). The terms "generate", "framework", and "develop", along with their corresponding word-forms, were used to appropriately limit the pool of publications in the unified set.

Inclusion and Exclusion. We focused on papers within the domain of HC-XAI. To be included, papers needed to present original work that introduced, applied, proposed, or evaluated human-centered methods or theories for explaining AI techniques, models, or systems. The initial filtering of papers (screening) was based on title, abstract and author keywords. Our paper selection process focused on original research articles that underwent peer-review and were written in English. Prior to evaluating the main content of each paper based on our inclusion criteria (eligibility), we manually excluded papers that comprised extended abstracts, doctoral consortium submissions, early career tracks, invited talks or tutorials, as well as all records which were not written in English.

PRISMA Guidelines. The flow diagram for the systematic literature review is shown in Figure 1, in which we illustrate different phases in agreement with the PRISMA guidelines.

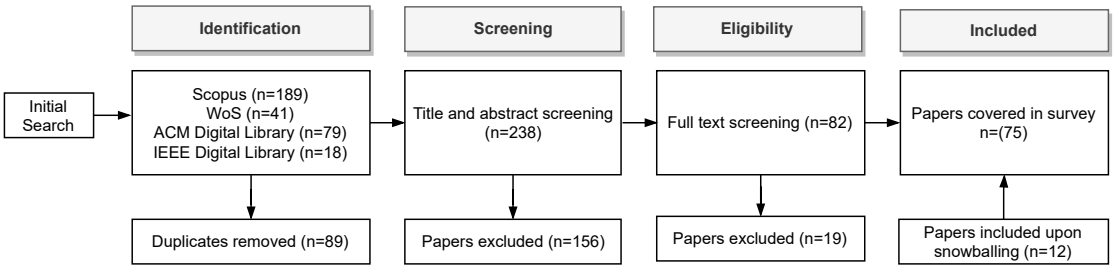

Figure 1: The flow diagram of the HC-XAI systematic literature review.

In the identification phase, we conducted an initial search across the selected digital libraries, using the predefined search query mentioned earlier. The results from this search yielded 18 papers from the IEEE digital library, 79 papers from the ACM digital library, 41 papers from WoS, and 189 papers from Scopus making a total of 327 papers. These relatively low numbers can be attributed to the limited research activity on HC-XAI.

In the screening phase, the relevant information (i.e., title, abstract, keywords, authors' names, journal name, source, publisher, and year of publication) from the identified records was exported to a Microsoft Excel spreadsheet. A total of 238 records were left after 89 duplicates were removed. These 238 papers were then subjected to screening based on the inclusion criterion. After the screening process, 156 papers were deemed ineligible and excluded from further consideration. The remaining 82 papers entered the eligibility phase, where each paper underwent a thorough analysis through full-text reading. In the eligibility phase, a decision was made on each paper, determining whether it would proceed to the final phase. Unfortunately, 19 papers did not adequately address to our research questions and were excluded, leaving a total of 63 papers for the final phase.

During the final inclusion phase, 12 additional papers were included through a snowballing process [31]: we identify and add other relevant publications manually from the bibliographies of the already selected manuscripts and relevant papers suggested by the peers of the authors, ensuring maximum coverage of the related subject areas. As a result, we had a set of 75 papers for synthesis and detailed meta-analysis.

Review Process. We conducted a thorough analysis of the final set of papers, following the taxonomy proposed by Guidotti et al. [24]. To delve into the specifics of this taxonomy, we kindly refer interested readers to Section 4 of Guidotti et al. survey [24]. In short, we examined and categorised contributions from multiple perspectives. The process began by reviewing the main content (excluding appendices and supplementary material) of each paper. Then, papers were categorised across four main dimensions: (1) the main contribution—whether the papers introduce, devise, or apply an approach, (2) the type of input data, (3) the specific type of task, and (4) the type of explanation.

4. Quantitative Analysis

Before addressing the research questions and related answers, we conducted a bibliometric analysis on the aggregated query results. This analysis allowed us to gain a comprehensive understanding of the HC-XAI research field.

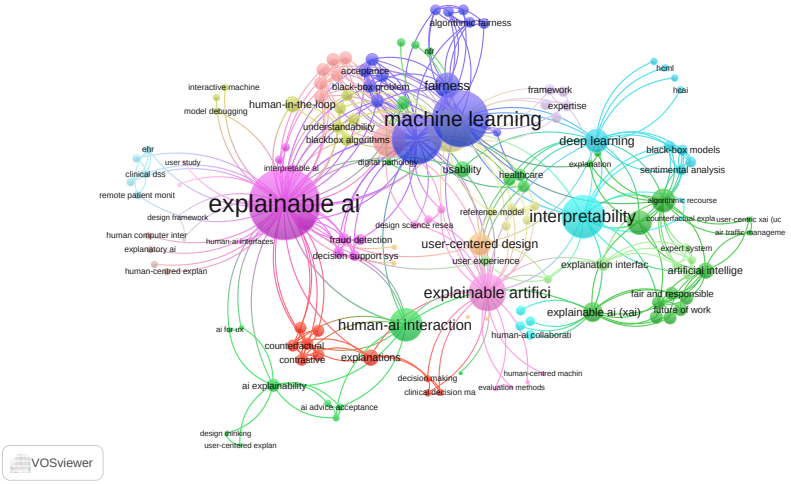

Figure 2: The map of author keywords for all papers under study.

Figure 2 depicts a graph (generated with the VOSViewer tool [32]) that presents the most common author keywords associated to the searched query. From the analysis, it can be inferred that HC-XAI is frequently explored within the context of human-AI interaction, as indicated by keywords such as "user-centered design", "user experience" and "human interaction". Additionally, the keywords "machine learning" (ML) and "interpretability" highlight the close connection between HC-XAI and XAI as a broader field.

4.1. Statistics of Included Papers

Figure 3 illustrates the growth of papers in the field of HC-XAI over the past years (from 2018 to mid-May 2023). The small number of papers published in 2018 (5) and 2019 (6), could be attributed to the relatively lower popularity of HC-XAI during those years (which coincide with the initial years of application of the GDPR) or to the limited usage of the related terms (i.e., "human-centered" and "user-centered") at that time. Anyway, there has been a consistent increase in the number of papers since 2020, with a very significant rise in 2022, even if we do not observe in HC-XAI the same exponential growth noticed by Adadi and Berrada [1] for XAI in general. Notice that, the decrease in 2023 is due to the fact that only 4.5 months are taken into account. All in all, the observed trend suggests a growing awareness of the importance and necessity of HC-XAI methods in recent years. Finally, regarding a comparison among papers that apply (blue line), evaluate (orange line), devise (green line), and other surveys and theories (red line) for HC-XAI has been shown in 3. There is a significant trend for devising methods

which seem to be gaining more and more attention.

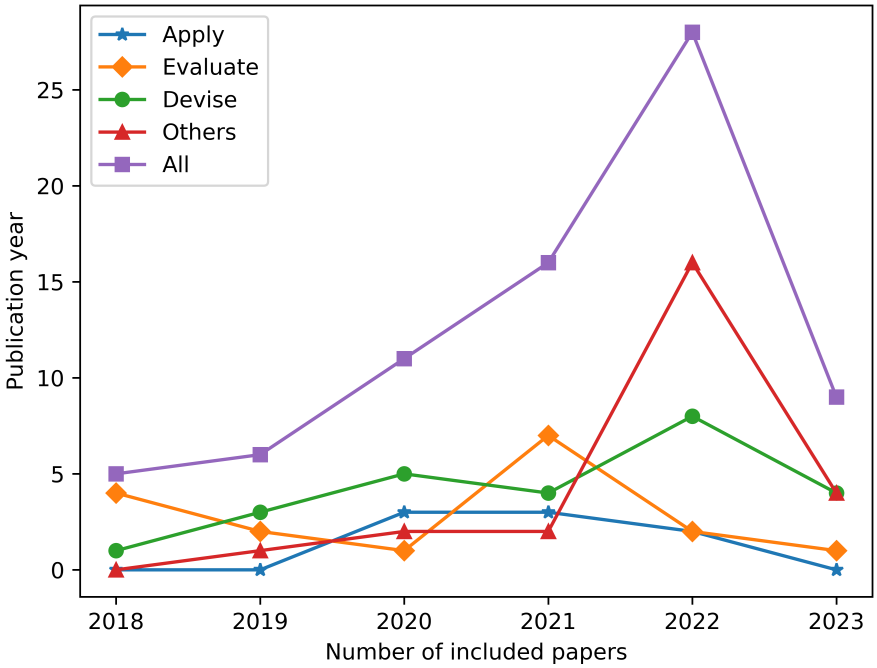

Figure 3: Distribution of papers per year that apply, evaluate, or devise an HC-XAI method (the magenta line aggregates all the papers including theories and surveys on HC-XAI).

4.2. Categorisation of Included Papers

The categorisation process considers various dimensions such as the type of data used, the specific task being addressed, the approach taken in contributing to the field (whether it involves application, evaluation, or development), and the specific type of explanation employed to clarify the underlying predictions or decisions made by the intelligent system.

4.2.1. Type of Data, Task, and Explanation

Different data types are taken as input for predictive models such as image, text, and tabular data. Our findings indicate that image data (8%) predominantly constitute the majority in heatmap explanations, tabular data (57%) for feature scores and counterfactual explanations while text input (6%) is primarily employed for textual explanations. The category "Other" (28%) is introduced to account for such data types as time series, graphs, or logical rules (ontology) that do not fit the predefined categories. Of course, it is important to note that these statistics may be influenced by our choice of publication venues, and the complexity of the task itself can also have an impact. Table 1 presents a quantitative analysis of various types of tasks, along with the related data types for those tasks, in the included papers and their corresponding bibliographical references. These are 49 papers out of 75 which solve any task in their contributing work. As shown in Table 1, a significant portion (51%) of the papers focus on classification models.

This can be attributed to the broad nature of classification tasks and the popularity of posthoc explanations, as explaining specific classification decisions serves as a compelling use case for outcome explanations. We observed that classification tasks were performed by using mainly tabular data, image or text, but also other data such as time series or graphs [8, 2, 33, 34, 35].

Table 1
Type of tasks solved by 49 out of 75 papers for HC-XAI.

Type of Task	Tabular data	Image	Text	Other data
Classification	[36, 37, 38, 39, 40, 16, 41, 42, 43, 44, 45, 46, 47, 48, 49]	[50, 51, 52, 53]	[54]	[8, 2, 33, 34, 35]
Regression	[55, 56, 57]			
Classification and Regression (both)	[58, 59, 60, 61, 62, 63]		[64]	[65, 63, 66, 67, 68, 22, 69]
Recommendation	[14, 70, 71, 72]			[73]
Others			QA [74]	Ontology [75]

The second largest category of papers (29%) addressed both classification and regression tasks concurrently in their studies. Notice that, papers which only deal with regression tasks consider only tabular data [55, 56, 57]. In addition, papers which account for recommendation tasks, they work with tabular [14, 70, 71, 72] and other graph-based data [73]. In other tasks, Jiao et al. [74] use textual data for investigating the task of question answering (QA) through the scenario-based design, while Shruthi et al. devise an explanation supported by ontological data [75].

On the other hand, we found out a diversity in the types of explanations under study. We categorised specific explanations in five main distinct categories which apply, devise, and evaluate (49 out of 75 papers), as outlined in Table 2. Notably, the most prevalent explanation type is the so-called hybrid approach which combines feature scores and counterfactual methods, which was found in 37% of the papers. This is followed by the use of counterfactuals alone (27%), feature scores alone (20%), and text highlighting (8%). Other explanation types include saliency maps, heatmaps, ontology, and rules, each accounting for 2% of the papers.

Table 2
Types of explanations in 49 out of 75 papers for HC-XAI.

Explanation Types	Included Papers
Feature Scores	[37, 41, 56, 43, 44, 45, 48, 34, 35, 60]
Counterfactuals	[36, 38, 52, 58, 40, 51, 16, 42, 70, 46, 47, 54, 49]
Hybrid	[8, 2, 62, 39, 65, 76, 61, 55, 33, 77, 78, 66, 67, 59, 68, 22, 69, 72]
Text Highlighting	[14, 74, 64, 71]
Others	Saliency maps [50], Heatmaps [52], Rules [73], Ontology [75]

Moreover, the analysis also highlights the dominance of approaches for explaining black-box models in current research on human-AI interaction. Local feature explanations, exemplified by LIME [9] and SHAP [10], are commonly employed to answer the "why" question. Counterfactual explanations, on the other hand, are used to tackle "what-if" or "why-not" questions. In the context of recommendation systems, a hybrid approach combining text highlighting, rules, and counterfactuals is also utilised, such as in a hybrid recommendation system [62]. Other

explanation types include text highlighting, saliency maps, heatmaps, rules, and an ontological representation of explanations. Text highlighting, as referenced in papers [14, 74, 64, 71], allows for emphasising specific textual elements to aid in understanding the explanation. Saliency maps, as employed in one of the papers [50], provide visual representations highlighting the most relevant areas or features. Heatmaps [52] offer a graphical depiction of the intensity or importance of different elements within the explanation. Rules, as discussed in another paper [73], provide structured statements or conditions to explain AI decisions. Lastly, the utilisation of an ontological representation [75] of explanations is also observed. These additional techniques contribute to the diverse range of approaches used for generating explanations in the included papers.

4.2.2. Type of Contribution and HC-XAI Methods.

The objective of this section is to provide insights to researchers and practitioners seeking appropriate contributions in the field of HC-XAI. In Table 3, we present the categorisation of papers that devise (introduce), apply, and evaluate the existing methods along with their specific contribution. It is worth noting that when a method devises an HC-XAI approach, it can be linked to multiple types of explanations, as there is no precondition of one-to-one correspondence between a method and an explanation type in the set of papers under study.

Table 3
Fundamental contributions of the 75 papers under study.

Type of Contribution	Included Papers
Theoretical Designs	[36, 79, 33, 59, 78, 80, 57, 68, 22, 81, 11]
User Studies and Interviews	[2, 62, 38, 37, 40, 51, 76, 55, 17, 77, 14, 42, 43, 70, 44, 45, 74, 54, 50, 67, 35, 49, 64, 82, 22, 69, 72]
Design Paradigms for Explanations	[8, 52, 58, 39, 53, 65, 76, 16, 41, 56, 46, 47, 48, 75, 78, 66, 34, 48, 60, 83, 17, 84, 71, 85]
Cognitive and Sociotechnical Theories	[86, 61, 87, 88, 15, 12, 89, 51, 90, 91, 92]
Highlighting HC-XAI Needs	[93, 94]
Systematic Literature Surveys	[13, 95, 96]

By examining the included papers, a categorisation of main contributions is presented in Table 3. The first group of papers provides theoretical guidelines for designing HC-XAI frameworks. The second group consists of user studies and interviews. We also refer to previous user studies conducted on ML systems [44] or explainable interfaces [17], which can be used for comparison or serve as templates for designing user studies [22]. The third group includes design paradigms for explanations. The fourth group deals with cognitive and socio-technical theories. For example, Miller [15] suggested building XAI on social sciences such as cognitive science and psychology. Additionally, XAI aims to assist users in developing mental models of AI systems [51]. The fifth group highlights the needs for HC-XAI [93, 94]. Finally, the sixth group includes systematic surveys [13, 95, 96].

5. Qualitative Analysis

In this section, we go in depth with answering, from a qualitative viewpoint, the three research questions that we posed in the introduction.

5.1. XAI Objectives Identified in Included Papers

RQ1 deals with frameworks, methods, theories and artefacts that support HC-XAI. First of all, we paid attention to generic concerns in the context of XAI (see Table 4).

Table 4
The 55 out of 75 papers which identify aims and objectives for HC-XAI.

Objective	Publication
Trust, transparency, and fairness	[44, 37, 52, 65, 55, 56, 44, 46, 74, 47, 49]
Improved User satisfaction	[8, 16, 41, 14, 70, 48, 54, 87, 93, 75, 35, 60, 92, 72, 71]
User experience	[2, 40, 76, 77, 78, 48, 17, 85]
User goals	[71]
Safety	[67]
Social needs	[86, 51, 61, 64, 15, 12, 89]
Aid user's future actions and promote informed decision-making	[38, 58, 42]
Improve human-AI collaboration and task performance	[62, 39, 73, 45, 74, 50, 66, 84, 97]

The primary studies highlighted multiple objectives for XAI, with many papers emphasising the general reasons and prerequisites for their implementation. Fostering user trust and enhancing transparency emerged as the most prominent goals among the commonly identified objectives. Another recurring theme observed in the studies revolved around catering to the user satisfaction as well as user needs, user experience, and social needs across diverse domains. This emphasis stems from the fact that different users (including domain experts, end-consumers, AI engineers, legislators...) possess unique motivations for seeking explanations. The purpose of XAI is to effectively assist each user in achieving their specific objectives by tailoring the explanations to their individual needs. User-centered approaches ensure that XAI accommodates properly diverse requirements and aspirations of users across various domains. We only found a few papers to aid user future actions and promoting informed decision-making. Other recurring objectives which require further attention are related to the improvement of the human-AI collaboration and thus increasing the task performance.

5.2. Interaction and Dialogue-based Explanation Methods

RQ2 investigated XAI techniques that are supported by a co-design perspective, i.e., they incorporate interaction or dialogue as a human-centered element and play a crucial role in enhancing transparency and understanding of AI systems. These techniques aim to bridge the gap between complex ML models and human users by enabling meaningful conversations and explanations. Through the integration of dialogue, XAI techniques foster a more inclusive and user-centered approach, ensuring that AI systems are not black boxes but rather tools that

empower individuals to make informed decisions. For example, by allowing users to engage in a dialogue with an AI system, Stepin et al. [98] facilitate a two-way exchange of information through an information-seeking dialogue, where the AI system provides explanations for its decisions and users can seek clarification or express their concerns.

In addition, Akula et al. [52] introduced a mind-based framework to enhance user's trust in counterfactual explanations for image data. The end user can interact to seek desired counterfactual outcomes through a dialogue-based interface. Cabour et al. [39] also highlighted the benefit of combining the technical development of XAI systems with a proper identification and interpretation of the user needs. They proposed an architecture to define the explanation space from a user-inspired perspective. This architecture mainly caters to five elements: the end-users' mental models, cognitive process, interface, the human-explainer agent, and the agent process. Suffian et al. [16] also remarked the need for human-involvement in the explanation generation process by customising explanations with user feedback. Their framework customises counterfactual explanations on demand. Finally, Zhu et al. [68] proposed an XAI framework for Designers, specifically tailored for game designers. Their human-centered approach facilitates game designers to co-create with AI techniques by focusing on specific users, their needs and tasks. Human-AI interaction and dialogue promote trust, accountability, and collaboration, as users gain insights into the decision-making process and can provide input and feedback.

5.3. Theoretical and Practical Aspects of Explanations

RQ3 investigated HC-XAI methods which are grounded on theoretical and practical evaluation approaches. Evaluation of explanations can be supported by functionally-grounded, human-grounded, or application-grounded approaches (see Table 5). We adhere to the categorisation of metrics proposed by Doshi-Velez and Kim [19].

Table 5
The papers devising grounded approaches for HC-XAI.

Functionally Grounded	**Human Grounded**	**Application Grounded**
loss minimisation [44], proxy measures and complexity check [58], interface for output [36], visualisation tool [41] [47], competency questions [75], [34, 60], user study [64]	user input [16], user study [2, 62, 38, 37, 40, 51, 76, 55, 17, 77, 14, 42, 43, 70, 44, 45, 74, 54, 50, 67, 35, 49, 64, 82, 22, 69, 72], interface [66], multiple user interfaces [48], co-design [22, 69], online survey [71]	user study [8, 52, 73, 68]

Functionally-grounded metrics do not require any human feedback. They are based on objective criteria that can be measured without human involvement. As a result, they measure the formal properties of the explainer, while regarding functional aspects of explanations (e.g., fidelity, accuracy, actionability, sparsity, or plausibility). *Human-grounded* metrics require direct human involvement for their measurement. They emphasise the human perspective and consider cognitive and psychological factors (e.g., satisfaction, persuasiveness, or novelty). These metrics gather feedback from users to assess the quality and usefulness of explanations. *Application-grounded* metrics focus on real-world applicability and impact. Their computation usually involves domain experts who are asked to assess how well the intelligent system performs

within specific application domains, considering practical constraints and requirements.

6. Research Opportunities

Insights from the systematic literature review carried out in the previous sections indicate a pressing need to develop human-centered explanation approaches in various fields of AI. Traditional methods of explaining AI decisions often fall short in meeting human requirements for transparency and understanding. However, a promising avenue that has gained considerable attention is the utilisation of human-perspective for explanation generation and evaluation.

Human-centered explanations play a critical role in promoting transparency, trust, and accountability in AI systems. Such explanations enable users to understand the reasoning behind automated decisions and empower them to make informed choices. However, existing research indicates a gap in actionable and human-centered explanations, what leads to some research opportunities for the exploration of human-centered approaches in the field of HC-XAI:

- **Bridging the explanation gap by handling user needs.** The systematic literature review reveals a persistent gap in providing actionable explanations that are meaningful and relevant to users. The explanations driven by user feedback offer a promising approach to bridge the gap. By incorporating user preferences and context, explanations can provide actionable insights that resonate with users and facilitate more informed decision-making.
- **Personalisation and Contextualisation.** Different users may have diverse preferences, values, and needs, and a one-size-fits-all explanation may not be sufficient. By involving users in the explanation generation process, explanations can adapt to individual user characteristics and provide insights that are tailored to their specific context.
- **Closing the Feedback Loop.** The literature review indicates a need to close the feedback loop between users and AI systems, enabling iterative improvements and accountability. This iterative feedback loop can enhance system performance, address biases and limitations, but also ensure the system evolves based on user needs and expectations.
- **Human-AI Collaboration and Co-creation.** Human-centred explanations promote human-AI collaboration and co-creation. By involving users in the generation and evaluation of explanations, we facilitate a meaningful dialogue between users and AI systems. Users can: provide feedback, ask questions, actively participate in refining and improving the system's behaviour.

7. Conclusion

In this work, we have presented a systematic literature review from which we can conclude that there is a need for more human-centered explanation approaches. Such HC-XAI approaches can enhance human understanding, enable ethical evaluations, and foster user engagement, thereby serving the "human-in-the-loop" cause. Moreover, building human-centered explanations means bridging the explanation gap, enhancing user agency, enabling personalisation and

contextualisation, addressing trust and ethical concerns, closing the feedback loop, and promoting human-AI collaboration. Consequently, we can bridge the gap between complex algorithms and human intuition, empowering individuals to make informed decisions, identifying biases, and actively taking part in the development of responsible AI technologies.

Even though we conducted a medium-scale survey, it was more feasible and practical than larger surveys, providing valuable insights into human-centered approaches, while requiring fewer resources, such as time and human resources. Despite its limitations, this survey can provide a starting point for further research and help determine the viability of pursuing larger-scale surveys. In conclusion, various factors can influence the effectiveness of explanations in human-centered perspectives. These factors include the specific explanation technique used, the characteristics of the dataset, and the nature of the task at hand.

Acknowledgments

Muhammad Suffian is a PhD researcher (Matricola N.309445). Ilia Stepin is an FPI researcher (grant number: PRE2019-090153). This research was funded by MCIN/AEI/10.13039/501100011033 (grants PID2021-123152OB-C21, TED2021-130295B-C33, and RED2022-134315-T), the Galician Ministry of Culture, Education, Professional Training, and University (grants ED431C2022/19 and ED431G2019/04). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE access 6 (2018) 52138–52160.

[2] D. Kim, Y. Song, S. Kim, S. Lee, Y. Wu, J. Shin, D. Lee, How should the results of artificial intelligence be explained to users?-research on consumer preferences in user-centered explainable artificial intelligence, Technological Forecasting and Social Change 188 (2023) 122343.

[3] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. D. Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion (2023) 101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148. doi:https://doi.org/10.1016/j.inffus.2023.101805.

[4] P. Voigt, A. Von dem Bussche, The EU general data protection regulation (GDPR), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (2017) 10–5555.

[5] Parliament and Council of the European Union, Proposal for laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2023. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

[6] T. Ha, Y. J. Sah, Y. Park, S. Lee, Examining the effects of power status of an explainable artificial intelligence system on users' perceptions, Behaviour & Information Technology 41 (2022) 946–958.

[7] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.
[8] X. He, Y. Hong, X. Zheng, Y. Zhang, What are the users' needs? design of a user-centered explainable artificial intelligence diagnostic system, International Journal of Human–Computer Interaction 39 (2023) 1519–1542.
[9] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).
[10] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.
[11] M. Ribera, A. Lapedriza, Can we do better explanations? a proposal of user-centered explainable AI, in: IUI workshops, volume 2327, 2019, p. 38.
[12] U. Ehsan, M. O. Riedl, Human-centered explainable AI: Towards a reflective sociotechnical approach, in: HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22, Springer, 2020, pp. 449–466.
[13] H. Chen, C. Gomez, C.-M. Huang, M. Unberath, Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review, npj Digital Medicine 5 (2022) 156.
[14] J. Graefe, S. Paden, D. Engelhardt, K. Bengler, Human centered explainability for intelligent vehicles–a user study, in: Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2022, pp. 297–306.
[15] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[16] M. Suffian, P. Graziani, J. M. Alonso, A. Bogliolo, FCE: Feedback based counterfactual explanations for explainable AI, IEEE Access 10 (2022) 72363–72372.
[17] Q. V. Liao, M. Pribić, J. Han, S. Miller, D. Sow, Question-driven design process for explainable AI user experiences, arXiv preprint arXiv:2104.03483 (2021).
[18] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, M. O. Riedl, Operationalizing human-centered perspectives in explainable AI, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–6.
[19] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[20] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021) 1–45.
[21] J. J. Ferreira, M. S. Monteiro, What are people doing about XAI user experience? a survey on AI explainability research and practice, in: Design, User Experience, and Usability. Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22, Springer, 2020, pp. 56–73.
[22] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable AI, in: Proceedings of the 2019 CHI conference on human factors in computing systems,

2019, pp. 1–15.
[23] Y. Rong, T. Leemann, T.-t. Nguyen, L. Fiedler, T. Seidel, G. Kasneci, E. Kasneci, Towards human-centered explainable AI: User studies for model explanations, arXiv preprint arXiv:2210.11584 (2022).
[24] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.
[25] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, C. Tan, Towards a science of human-AI decision making: a survey of empirical studies, arXiv preprint arXiv:2112.11471 (2021).
[26] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–18.
[27] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (2020) 18.
[28] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, arXiv preprint arXiv:1902.01876 (2019).
[29] M. Chromik, M. Schuessler, A taxonomy for human subject evaluation of black-box explanations in XAI, ExSS-ATEC@ iui 1 (2020).
[30] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, t. PRISMA Group*, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Annals of internal medicine 151 (2009) 264–269.
[31] R. Streeton, M. Cooke, J. Campbell, Researching the researchers: Using a snowballing technique, Nurse researcher 12 (2004) 35–47.
[32] N. Van Eck, L. Waltman, Software survey: VOSviewer, a computer program for bibliometric mapping, scientometrics 84 (2010) 523–538.
[33] S. Hanses, J. Wang, How do users interact with AI features in the workplace? understanding the AI feature user journey in enterprise, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–7.
[34] E. Veitch, O. A. Alsos, Human-centered explainable artificial intelligence for marine autonomous surface vehicles, Journal of Marine Science and Engineering 9 (2021) 1227.
[35] M. Riveiro, S. Thill, "that's (not) the output i expected!" on the role of end user expectations in creating explanations of AI systems, Artificial Intelligence 298 (2021) 103507.
[36] M. Afzaal, A. Zia, J. Nouri, U. Fors, Informative feedback and explainable AI-based recommendations to support students' self-regulation, Technology, Knowledge and Learning (2023) 1–24.
[37] V. Swamy, S. Du, M. Marras, T. Kaser, Trusting the explainers: Teacher validation of explainable artificial intelligence for course design, in: LAK23: 13th International Learning Analytics and Knowledge Conference, 2023, pp. 345–356.
[38] A. Bhattacharya, J. Ooge, G. Stiglic, K. Verbert, Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support What-If explorations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 204–219.
[39] G. Cabour, A. Morales-Forero, É. Ledoux, S. Bassetto, An explanation space to align user studies with the technical development of explainable AI, AI & SOCIETY (2022) 1–19.

[40] C. Bove, M.-J. Lesot, C. A. Tijus, M. Detyniecki, Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 188–203.
[41] F. Cheng, Y. Ming, H. Qu, Dece: Decision explorer with counterfactual explanations for machine learning models, IEEE Transactions on Visualization and Computer Graphics 27 (2020) 1438–1447.
[42] T. Susnjak, G. S. Ramaswami, A. Mathrani, Learning analytics dashboard: a tool for providing actionable insights to learners, International Journal of Educational Technology in Higher Education 19 (2022) 12.
[43] N. Mollaei, C. Fujao, L. Silva, J. Rodrigues, C. Cepeda, H. Gamboa, Human-centered explainable artificial intelligence: automotive occupational health protection profiles in prevention musculoskeletal symptoms, International Journal of Environmental Research and Public Health 19 (2022) 9552.
[44] M. Dikmen, C. Burns, 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 (2022) 102792.
[45] W.-J. She, K. Senoo, H. Iwakoshi, N. Kuwahara, P. Siriaraya, Af'fective design: Supporting atrial fibrillation post-treatment with explainable AI, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 22–25.
[46] M. Suffian, A. Bogliolo, Investigation and mitigation of bias in explainable AI, in: CEUR WORKSHOP PROCEEDINGS, volume 3319, AIxIA, 2022, pp. 89–94.
[47] N. Spreitzer, H. Haned, I. van der Linden, Evaluating the practicality of counterfactual explanations, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022.
[48] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. Van Den Bosch, Human-centered XAI: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684.
[49] Y. Nakao, S. Stumpf, S. Ahmed, A. Naseer, L. Strappelli, Toward involving end-users in interactive human-in-the-loop AI fairness, ACM Transactions on Interactive Intelligent Systems (TiiS) 12 (2022) 1–30.
[50] Y. Gao, T. S. Sun, L. Zhao, S. R. Hong, Aligning eyes between humans and deep neural network through interactive attention alignment, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–28.
[51] K. Z. Gajos, L. Mamykina, Do people engage cognitively with AI? impact of ai assistance on incidental learning, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 794–806.
[52] A. R. Akula, K. Wang, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Chai, S.-C. Zhu, CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models, Iscience 25 (2022) 103581.
[53] J. Labaien Soto, E. Zugasti Uriguen, X. De Carlos Garcia, Real-time, model-agnostic and user-driven counterfactual explanations using autoencoders, Applied Sciences 13 (2023) 2912.
[54] M. Riveiro, S. Thill, The challenges of providing explanations of AI systems when they

do not behave like users expect, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 110–120.
[55] H. Ma, K. McAreavey, R. McConville, W. Liu, Explainable AI for non-experts: Energy tariff forecasting, in: 2022 27th International Conference on Automation and Computing (ICAC), IEEE, 2022, pp. 1–6.
[56] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, M. Detyniecki, Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users, in: 27th international conference on intelligent user interfaces, 2022, pp. 807–819.
[57] B. Hayes, M. Moniz, Trustworthy human-centered automation through explainable AI and high-fidelity simulation, in: Advances in Simulation and Digital Human Modeling: Proceedings of the AHFE 2020 Virtual Conferences on Human Factors and Simulation, and Digital Human Modeling and Applied Optimization, July 16-20, 2020, USA, Springer, 2021, pp. 3–9.
[58] M. Förster, P. Hühn, M. Klier, K. Kluge, User-centric explainable AI: design and evaluation of an approach to generate coherent counterfactual explanations for structured data, Journal of Decision Systems (2022) 1–32.
[59] D. Cirqueira, M. Helfert, M. Bezbradica, Towards design principles for user-centric explainable AI in fraud detection, in: Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Springer, 2021, pp. 21–40.
[60] A. A. Shrotri, N. Narodytska, A. Ignatiev, K. S. Meel, J. Marques-Silva, M. Y. Vardi, Constraint-driven explanations for black-box ML models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 8304–8314.
[61] H. Baniecki, D. Parzych, P. Biecek, The grammar of interactive explanatory model analysis, Data Mining and Knowledge Discovery (2023) 1–37.
[62] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, M. Gombolay, Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human-agent interaction, International Journal of Human–Computer Interaction 39 (2023) 1390–1404.
[63] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: informing design practices for explainable AI user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15.
[64] J. E. Sales, A. Freitas, S. Handschuh, A user-centred analysis of explanations for a multicomponent semantic parser, in: Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25, Springer, 2020, pp. 37–44.
[65] F. Sovrano, F. Vitali, Explanatory artificial intelligence (YAI): human-centered explanations of explainable AI and complex data, Data Mining and Knowledge Discovery (2022) 1–28.
[66] C. Panigutti, A. Beretta, D. Fadda, F. Giannotti, D. Pedreschi, A. Perotti, S. Rinzivillo, Co-design of human-centered, explainable AI for clinical decision support, ACM Transactions on Interactive Intelligent Systems (2023).
[67] D. H. Kim, E. Hoque, M. Agrawala, Answering questions about charts and generating visual explanations, in: Proceedings of the 2020 CHI conference on human factors in

computing systems, 2020, pp. 1–13.
[68] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. M. Youngblood, Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation, in: 2018 IEEE conference on computational intelligence and games (CIG), IEEE, 2018, pp. 1–8.
[69] X. Xu, A. Yu, T. R. Jonker, K. Todi, F. Lu, X. Qian, J. M. Evangelista Belo, T. Wang, M. Li, A. Mun, et al., XAIR: A framework of explainable AI in augmented reality, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–30.
[70] R. Shang, K. K. Feng, C. Shah, Why am i not seeing it? understanding users' needs for counterfactual explanations in everyday recommendations, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1330–1340.
[71] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, User preferences for hybrid explanations, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, pp. 84–88.
[72] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, Personalized explanations for hybrid recommender systems, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 379–390.
[73] P. Qian, V. Unhelkar, Evaluating the role of interactivity on improving transparency in autonomous agents, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, 2022, pp. 1083–1091.
[74] J. Sun, Q. V. Liao, M. Muller, M. Agarwal, S. Houde, K. Talamadupula, J. D. Weisz, Investigating explainability of generative AI for code through scenario-based design, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 212–228.
[75] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation ontology: a model of explanations for user-centered AI, in: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II, Springer, 2020, pp. 228–243.
[76] S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, O. Russakovsky, Hive: evaluating the human interpretability of visual explanations, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII, Springer, 2022, pp. 280–298.
[77] J. Novak, T. Maljur, K. Drenska, Transferring AI explainability to user-centered explanations of complex COVID-19 information, in: HCI International 2022–Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Springer, 2022, pp. 441–460.
[78] Q. V. Liao, D. Gruen, S. Miller, Questioning the AI: informing design practices for explainable AI user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15.
[79] L. Wiebelitz, P. Schmid, T. Maier, M. Volkwein, Designing user-friendly medical AI applications-methodical development of user-centered design guidelines, in: 2022 IEEE International Conference on Digital Health (ICDH), IEEE, 2022, pp. 23–28.
[80] S. Naveed, J. Ziegler, Featuristic: An interactive hybrid system for generating explainable recommendations-beyond system accuracy., in: IntRS@ RecSys, 2020, pp. 14–25.
[81] A. Kirsch, Explain to whom? putting the user in the center of explainable AI, in:

Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017), 2017.

[82] L. Yang, H. Wang, L. A. Deleris, What does it mean to explain? a user-centered study on AI explainability, in: Artificial Intelligence in HCI: Second International Conference, AI-HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Springer, 2021, pp. 107–121.

[83] L. Sanneman, J. A. Shah, A situation awareness-based framework for design and evaluation of explainable AI, in: Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2, Springer, 2020, pp. 94–110.

[84] M. El-Assady, C. Moruzzi, Which biases and reasoning pitfalls do explanations trigger? decomposing communication processes in human–AI interaction, IEEE Computer Graphics and Applications 42 (2022) 11–23.

[85] S. F. Jentzsch, S. Höhn, N. Hochgeschwender, Conversational interfaces for explainable AI: a human-centred approach, in: Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1, Springer, 2019, pp. 77–92.

[86] U. Ehsan, K. Saha, M. De Choudhury, M. O. Riedl, Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–32.

[87] L. Sanneman, J. A. Shah, An empirical study of reward explanations with human-robot interaction applications, IEEE Robotics and Automation Letters 7 (2022) 8956–8963.

[88] M. Ghajargar, J. Bardzell, Making AI understandable by making it tangible: Exploring the design space with ten concept cards, in: Proceedings of the 34th Australian Conference on Human-Computer Interaction, 2022, pp. 74–80.

[89] L. Capone, M. Bertolaso, et al., A philosophical approach for a human-centered explainable AI, in: XAI.it@AIxIA, 2020, pp. 80–86.

[90] A. Kasirzadeh, A. Smart, The use and misuse of counterfactuals in ethical machine learning, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 228–236.

[91] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, J. D. Weisz, Expanding explainability: Towards social transparency in AI systems, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–19.

[92] G. Margetis, S. Ntoa, M. Antona, C. Stephanidis, Human-centered design of artificial intelligence, Handbook of human factors and ergonomics (2021) 1085–1106.

[93] U. Schmid, B. Wrede, What is missing in XAI so far? an interdisciplinary perspective, KI-Künstliche Intelligenz (2022) 1–13.

[94] C. M. Navarro, G. Kanellos, T. Gottron, Desiderata for explainable AI in statistical production systems of the european central bank, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I, Springer, 2022, pp. 575–590.

[95] C. T. Okolo, N. Dell, A. Vashistha, Making AI explainable in the global south: A systematic review, in: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies

(COMPASS), 2022, pp. 439–452.
[96] A. Bertrand, R. Belloum, J. R. Eagan, W. Maxwell, How cognitive biases affect XAI-assisted decision-making: A systematic review, in: Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society, 2022, pp. 78–91.
[97] M. Suffian, M. Y. Khan, A. Bogliolo, Towards human cognition level-based experiment design for counterfactual explanations, in: 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2022, pp. 1–5. doi:10.1109/MAJICC56935.2022.9994203.
[98] I. Stepin, K. Budzynska, A. Catala, M. Pereira-Fariña, J. M. Alonso-Moral, Information-seeking dialogue for explainable artificial intelligence: Modelling and analytics, Argument & Computation (2023).