Article:
HUGHES, NATHAN, JIA, YAN, Sujan, Mark Alexander et al. (2026) Design Principles for Human-Centred Explainable AI:A Scoping Review. ACM Transactions on Interactive Intelligent Systems.

https://doi.org/10.1145/3771720

PDF Download
3771720.pdf
10 February 2026
Total Citations: 0
Total Downloads: 412

Published: 09 January 2026
Online AM: 14 October 2025
Accepted: 01 October 2025
Revised: 16 September 2025
Received: 03 June 2024

RESEARCH-ARTICLE

Design Principles for Human-Centred Explainable AI: A Scoping Review

NATHAN GERARD J HUGHES, University of York, York, North Yorkshire, U.K.

YAN JIA, University of York, York, North Yorkshire, U.K.

MARK ALEXANDER SUJAN, University of York, York, North Yorkshire, U.K.

TOM LAWTON, Bradford Institute for Health Research, Bradford, West Yorkshire, U.K.

IBRAHIM HABLI, University of York, York, North Yorkshire, U.K.

JOHN A MCDERMID, University of York, York, North Yorkshire, U.K.

Open Access Support provided by:

University of York

Bradford Institute for Health Research

ACM Transactions on Interactive Intelligent Systems, Volume 16, Issue 1 (March 2026)
https://doi.org/10.1145/3771720
EISSN: 2160-6463

Design Principles for Human-Centred Explainable AI: A Scoping Review

NATHAN HUGHES, YAN JIA, and MARK SUJAN, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland
TOM LAWTON, Improvement Academy, Bradford Institute for Health Research, Bradford, United Kingdom of Great Britain and Northern Ireland
IBRAHIM HABLI and JOHN MCDERMID, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland

The field of Human-Centred Explainable AI (HCXAI) has been rapidly expanding. In turn, there has been an increase in the number of papers suggesting design principles for HCXAI. However, it is unclear the extent to which design requirements overlap between papers, and in turn what the field overall considers to be HCXAI design requirements. To overcome this, this study analysed the state of the field via a scoping review of papers suggesting HCXAI design requirements, and a Content Analysis of the extracted principles. A total of 330 design principles were identified from 35 papers, which were subsequently categorised into 43 codes and grouped into 4 main areas of focus. Based on these findings, we propose a definition of HCXAI which identifies HCXAI as a design process rather than an XAI technique. Finally, an overview of the current state of HCXAI is presented, as well as areas where further research is required.

CCS Concepts: • **Human-centered computing → Interaction design theory, concepts and paradigms**; **HCI theory, concepts and models**;

Additional Key Words and Phrases: Human-centred XAI, User-centric, Explainable AI, Scoping Review, Content Analysis

ACM Reference format:
Nathan Hughes, Yan Jia, Mark Sujan, Tom Lawton, Ibrahim Habli, and John McDermid. 2026. Design Principles for Human-Centred Explainable AI: A Scoping Review. *ACM Trans. Interact. Intell. Syst.* 16, 1, Article 4 (January 2026), 26 pages.
https://doi.org/10.1145/3771720

This project was funded by Lloyd's Register Foundation and University of York through Assuring Autonomy International Programme (Project ref 06/22/04) with continued support from the Centre for Assuring Autonomy.

Authors' Contact Information: Nathan Hughes (corresponding author), Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland; e-mail: nathan.hughes@york.ac.uk; Yan Jia, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland; e-mail: yan.jia@york.ac.uk; Mark Sujan, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland; e-mail: mark.sujan@york.ac.uk; Tom Lawton, Improvement Academy, Bradford Institute for Health Research, Bradford, United Kingdom of Great Britain and Northern Ireland; e-mail: tom.lawton@bthft.nhs.uk; Ibrahim Habli, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland; e-mail: ibrahim.habli@york.ac.uk; John McDermid, Computer Science, University of York, York, United Kingdom of Great Britain and Northern Ireland; e-mail: john.mcdermid@york.ac.uk.

ACM 2160-6463/2026/1-ART4
https://doi.org/10.1145/3771720

1 Introduction

Tools using AI have become increasingly complex and able to perform a wider range of tasks. Consequently, there has been a growing interest in designing AI tools that will be understood by the people who will interact with them. At the same time, there has been an increased demand for transparency from such tools, both in terms of the human–AI interactions that occur and the outputs given by the AI (e.g., [17]). This has led to a focus on creating **Human-Centred Explainable AI (HCXAI)**, which does not have a set definition, but generally refers to considering factors that influence human decision-making and communication during the design and development process [25]. HCXAI has been studied in a variety of applications, such as healthcare [25], fraud detection [14] and data privacy disclosure [10], where general HCXAI design principles are commonly derived to summarise findings. These design principles aim to help future designers and researchers build AI systems, by providing transferable guidance that could be applied to similar applications.

The growing interest in HCXAI has resulted in a substantial rise in papers proposing design principles for the field. However, many papers present principles derived from empirical studies (e.g., [16, 20, 23, 36, 41]), and discuss them in isolation, without comparison with existing principles in the literature. This opens up the possibility that, instead of advancing the field of HCXAI by providing new insights, suggested principles may have already been noted and presented previously. It therefore remains unclear to what extent the proposed HCXAI design principles overlap, or what principles the field as a whole is currently suggesting. In turn, this makes it difficult to understand what the field is currently prioritising within HCXAI, and identify potential gaps for future research.

To overcome these issues, several meta-reviews have been conducted to summarise the HCXAI design principles currently proposed, such as [13, 49] and [14]. However, as the field continues to expand rapidly, there is a need for an updated scoping review to identify any new design principles, areas of research priority and potential gaps for future research. Further, existing meta-reviews typically take the form of scoping or systematic literature reviews, followed by a narrative summary that concludes with a condensed list of design principles. However, these studies often lack clarity regarding how the authors analysed the papers or combined the data to conclude the condensed principles. Accompanying a scoping review with other qualitative research methods that treat design principles as data, such as Content Analysis [29], can overcome these limitations. This approach increases the transparency of data analysis and enables a more comprehensive exploration of the state of the HCXAI field, its evolving design principles and its current priorities.

Therefore, in this study, we performed a scoping review of current HCXAI design principles and analysed them via a Content Analysis, to answer the following research questions:

(1) What HCXAI design principles currently exist and what methods are used to identify them?
(2) What insights about the field of HCXAI can be gained from these design principles?

2 Background

With the advance of AI, especially deep learning, AI systems are able to perform increasingly complex tasks that previously needed to be handled by humans. This evolution has ushered in a new era of human–AI collaboration, where critical decisions can now be partly supported by AI. Communicating AI decisions and reasoning to humans is therefore important to ensure informed and safe decisions are made. However, without understanding how humans make decisions during a task, and how they communicate their actions to other humans involved in the task, it is difficult to design human–AI interactions that enable effective explanations for the given context.

As outlined in [25], the way in which AI explanations are designed and presented should be user-centred, by taking into account the factors which influence human decision-making and communication. This user-centred focus underpins the concept of HCXAI—an approach to designing human–AI interactions involving explanations of outputs/behaviours, whilst also considering elements of the wider system. In doing so, the explanations an AI system presents to a user can be more meaningful and relevant, both to the context and tasks being performed.

As it is important for AI explanations to be human-centred, there has been a growing interest in understanding how this can be achieved. An increasing body of knowledge of important factors to consider when designing interactions with AI systems has emerged aimed at producing meaningful and relevant explanations (i.e., HCXAI) [39]. For example, several studies have suggested HCXAI design principles either by conducting empirical studies with users directly (e.g., [2, 7, 28]) or from conducting literature reviews synthesising such studies (e.g., [13, 14, 46]). The focus of these literature reviews and the principles they suggest varies. For example, some provide general guidance for developing AI systems with human-centred explanations, such as [21] who conducted a systematic review of **Explainable AI (XAI)** from an end user's perspective, and related these to the five elements of HCXAI identified by Laato et al. [31]—trust, transparency, understandability, usability and fairness. The findings formed guidelines for designing AI systems with human-centred explanations, applicable to a variety of domains, such as digital services and education rather than focused solely on 'mission-critical systems.' Similarly, Muelle et al. [39] provided insights based on 'human-centered research on explanation in AI systems' summarised as 14 design principles for AI developers to refer to during development. Other papers apply findings from other research domains such as psychology to understand how they may benefit the development of AI systems. For example, Bertrand et al. [6] discuss how cognitive biases may influence decision-making with XAI systems, whilst Wang et al. [49] similarly highlight how understanding human cognition can help to develop XAI systems that support human reasoning.

Overall, findings in the literature are commonly summarised as a list of design principles for developers and researchers interested in HCXAI to use in future work. However, with a high volume of papers suggesting HCXAI design principles, especially for decision-support systems in healthcare (e.g., [11, 34, 41, 42]), it becomes challenging to effectively synthesise the research. Suggested principles show varying levels of overlap to previous papers, making it difficult to consolidate the findings. Gu et al. [20], for example, identify the traceability of explanations as an important feature of HCXAI, which was similarly concluded by Sokol and Flach [47]. Similarly, He et al. [22] identify the need to provide tailored explanations for the particular user and their circumstance, as does Miller [38]. However, the extent to which papers relate their design principles to previously existing work is variable; He et al. [22], for example, references Miller [38], whilst Gu et al. [20] does not reference Sokol and Flach [47]. Therefore, it is unclear the extent to which previous work in the field is incorporated into newer publications, and as such makes it difficult to assess how cohesive the suggested HCXAI design principles are overall. This ambiguity further complicates efforts to determine whether the field is generating genuinely new insights or largely reiterating existing principles without additional nuance.

To address this, meta-reviews have been conducted to synthesise studies and identify the current key HCXAI design principles. For example, Chromik and Butz [13] conducted a scoping review of 91 papers on human–XAI interactions, which summarised the results as 4 design principles for interactive Explainable User Interfaces (XUI). Similarly, Laato et al. [31] reviewed 25 papers on how to explain AI systems to end users, and suggested 16 design recommendations. From these meta-reviews, it is possible to observe the common areas of interest within the field of HCXAI, and the current types of design principles being proposed. Whilst these meta-reviews have provided useful summaries of HCXAI, e.g. [13] conducted in 2021 and [31] conducted in 2022, the field has continued to expand rapidly in recent years. Therefore, there is a need for an updated review to synthesise new design requirements and assess the current state of the field.

Furthermore, the current approach to conducting meta-reviews typically takes the form of a narrative summary, followed by a list of design principles. It is not always clear how the authors summarised the papers into the list of principles, or what methodology was used to analyse them. For example, Chromik and Butz [13] provide an in-depth explanation for how papers were found, but provide little information on how papers were then analysed once collected. Similarly, Laato et al. [31] explain the steps taken to find papers systematically, as well as how principles were extracted and analysed via axial coding, but the resulting themes and recommendations are linked only to the source papers, without showing how the specific principles within the papers helped to derive them. In both cases, it is therefore unclear how the analyses of the selected texts were conducted; for example, how did the authors analyse the selected texts? How many times were the selected texts analysed (i.e., how many passes of the data were performed, and by whom)? And how do the principles contained within the selected texts relate to the design principles the authors conclude with? Without access to the original principles used as the dataset, it is difficult to see how each principle maps to the recommendations given, reducing the transparency and traceability of the work undertaken. Both are important for result dissemination, as transparency in methodology aids the reader in how to interpret the conclusions of a meta-review, as they can then trace how individual papers/insights map to the final meta-review summary.

For these reasons, scoping reviews can be complemented with qualitative research methods to chart data with further transparency [4]. One such charting method is Content Analysis, which provides a way to combine large amounts of qualitative data into groups/codes that summarise what types of information and concepts exist within the dataset [29]. By treating HCXAI design principles as data, it is possible to highlight the common areas of interest within the field, which can reveal what the field has prioritised for study, and by extension, areas that have received less attention. From here, what is currently considered to be HCXAI design principles can be learnt, as well as areas of potential future research. Therefore, this study conducted a scoping review with Content Analysis to synthesise recent work in the field, with a secondary emphasis on transparent and traceable methodological reporting. The results provide a consolidated overview of the current state of HCXAI.

3 Method

The analysis in this study took place in two steps: a scoping review of the literature to extract HCXAI design principles, followed by a Content Analysis to summarise the principles.

3.1 Paper Collection

A scoping review was conducted to collect papers on general design principles for HCXAI, as these are useful for identifying knowledge gaps in the literature [40]. In a scoping review, it is possible to broadly map the relevant literature of a field by applying qualitative analysis to the collected papers [4]. As we are interested in going beyond a narrative description of the existing literature, a Content Analysis was selected for paper analysis [29]. For reporting the methodology and results, we followed the PRISMA extension guidance for scoping reviews [48]. The steps taken to identify papers eligible for inclusion are shown in Figure 1.

To find the initial papers, the ACM Digital Library and Google Scholar were searched. Eligible papers were those dating anytime before November 2023 and written in English. For the ACM Digital Library, the following search term was used: [[All: “human centred”] OR [All: “human centric”] OR [All: “user centred”] OR [All: “user centric”]] AND [[All: “explainable AI”] OR [All: “XAI”]] AND [[All: “design requirements”] OR [All: “design principles”]]. For Google Scholar, the following search term was used: [“Human” or “User”] and [“Centred” or “Centered” or “Centric”] and [“Explainable” or “Explanation”] and [“AI” or “XAI”] and [“Design”] and [“Principles” or

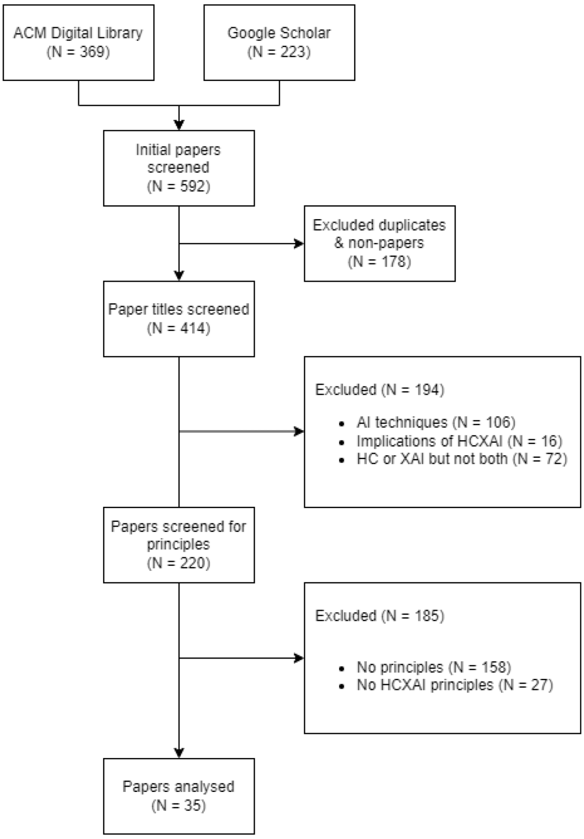
ACM Digital Library
(N = 369)
Google Scholar
(N = 223)
Initial papers screened
(N = 592)
Excluded duplicates & non-papers
(N = 178)
Paper titles screened
(N = 414)
Excluded (N = 194)
• AI techniques (N = 106)
• Implications of HCXAI (N = 16)
• HC or XAI but not both (N = 72)
Papers screened for principles
(N = 220)
Excluded (N = 185)
• No principles (N = 158)
• No HCXAI principles (N = 27)
Papers analysed
(N = 35)

Fig. 1. An overview of the paper screening process.

"Requirements"]. Two search strings were used due to differences in how search strings are written between websites, however care was taken to keep them as similar as possible. From this search, papers specifically discussing the intersection of human/user-centred design and explainable AI (i.e., HCXAI-focused papers) were collected to be included in the review.

The screening process was conducted by the first author, after deliberation with the co-authors to create the inclusion/exclusion criteria. The initial search resulted in 369 papers from ACM and 223 from Google Scholar. After removing conference proceeding summary documents, citations, books and duplicates, this list was reduced to 240 and 208, respectively. Duplicates across the sources were then removed, leaving 414 papers. The titles of the papers were then screened to find those specific to HCXAI. Papers that focused on the introduction of AI techniques (such as reinforcement learning) were removed, leaving 308 papers. Papers discussing the implications/benefits of creating HCXAI were then removed (such as the benefits for fairness), leaving 292 papers. Papers that discussed XAI but without a human-centred focus, or vice versa, were then removed, leaving 220 papers. The final 220 papers were then read by the first author to assess if they provided HCXAI design principles, either in the prose of the discussion or as a set of bulletpoints/tables. Removing papers that did not provide such principles left 62 papers, of which 35 were specific to HCXAI rather than solely to XAI or human-centred design. At this step, papers were distinguished based on whether they explicitly defined their principles as relating to HCXAI, rather than 'user-centred design' or 'explainable AI' in isolation. A full list of the papers selected for analysis can be found in the Supplementary Materials.

3.2 Paper Analysis

Alongside the number of principles derived, the following variables were extracted to provide a description of the papers used for analysis: how the principles were created (i.e., from an empirical study or literature review), what domain they relate to (e.g., healthcare), the expected users considered (e.g., non-domain experts) and whether the principles were designed to be domain-specific or generalisable across other domains/applications.

A total of 246 individual HCXAI design principles were extracted from the included papers by the first author. This was done by reading each paper for highlights of results; many papers provided bulletpoint lists or tables of key findings and design principles, however some papers instead contained summary paragraphs in the discussion. In the latter case, summary sentences were extracted and treated the same as those originally from bulletpoints. Some papers provided a broad range of design principles where only some were explicitly stated to relate to human-centred explanations. For example, Long et al. [36] reported 17 design principles ranging from technology design (e.g., modularity) to research design (e.g., informed consent), where only one principle was related to explainability. In these cases, only the principles specific to HCXAI were extracted. Furthermore, sometimes principles contained more than one principle; for example, the principle, '[...] explanation should be timely and adapted to the expertise of the stakeholder concerned' from [19] contains the two concepts of 'timely' and 'adapted'. To improve clarity and simplify the analysis, such multi-component principles were split into individual principles. This process yielded a total of 330 principles.

To analyse the 330 principles found in the literature, an inductive Content Analysis [29] was performed by the first author. Overall, the content analysis included four major steps, with iterations at each step. Specifically, the initial pass involved grouping the principles into loose categories based on what aspect of HCXAI they were referring to. For example, a number of principles made reference to a need to understand the context of where the AI system is to be deployed, and so the code, 'Consider the Wider Context' was created. Doing so generated 28 codes. The coding scheme and their assigned principles were then discussed with the second author to reach consensus. This led to a refinement in the codes to ensure they best reflected the data, which were implemented in a second pass of the data by the first author.

As a part of this process, many categories were found to contain a high volume of principles, and were therefore split into smaller codes to improve their individual definitions, resulting in 43 codes. This allowed for a granular analysis; however, it made it difficult to understand the entire dataset cohesively. To strike a balance between specificity of codes and interpretability overall, the codes were then analysed and grouped into a coding scheme consisting of hierarchical groups: codes, meta-codes and areas of focus. Meta-codes grouped similar codes together: for example, the codes, 'Is Explanation Needed' and 'Goal of Explanation' both refer to the reasons why a designer would create an AI system with explanation capabilities. Consequently, these both fall under the meta-code, 'Explain Reason for Providing Explanation'. Areas of focus reflect what aspect of HCXAI the meta-codes referred to, and were found by identifying similarities, patterns and relationships between meta-codes. This resulted in four areas of focus: the Human-Centred in HCXAI (54 principles), Design Aspects of HCXAI (51), the XAI in HCXAI (137), and Characteristics of HCXAI (88), as shown in Figure 2.

A third pass was then conducted with feedback from all co-authors to ensure the hierarchical groups were applied consistently, which resulted in 43 codes sorted into 10 meta-codes and 4 areas of focus. However, as all previous analyses were completed with co-author involvement, it was

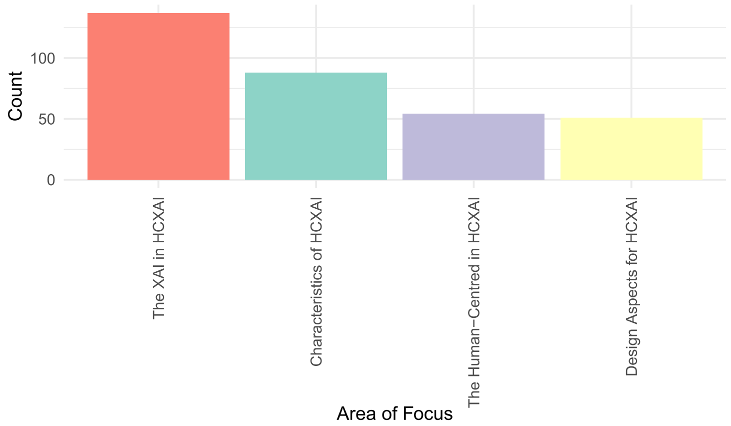

Fig. 2. A histogram of the total code counts for the areas of focus.

important to assess the reliability of the generated codes via the use of an independent second coder. Therefore, at the end an external second coder with expertise in qualitative data analysis for Human–Computer Interaction was included. The coder was blind from the previous analysis, and was given the 330 principles and the codebook, which contained a list and a description of each of the 43 codes. They were instructed to assign one code per principle, and results were compared to the first author's coding. To assess inter-rater reliability, Cohen's Kappa was calculated [15]. The threshold for the Kappa statistic was set as above 0.61, as this places it within the 'substantial' agreement category [32]. The calculated Kappa was found to be 0.66. On reflection, many disagreements related to the fact that the second coder was experienced with Human–Computer Interaction broadly rather than explainability algorithms—as such, there were misalignments for codes such as what constituted a type of feature importance. Other disagreements were related to similarities between categories, such as 'Why This Explanation' vs. 'General System Explanation'. The external coding process provided a useful reflection on the implicit assumptions made by the authors based on their familiarity with the topic. Disagreements were addressed by updating the wording of the code definitions to be clearer. A full list of the extracted design principles and the codes they fall under can be found in the Supplementary Materials.

4 Results

An overview of the papers collected for analysis is presented in Table 1. This shows the number of principles derived, how the principles were created, what domain they relate to, the expected types of users considered and whether the principles are designed to be generic or domain-specific.

As can be seen, many papers originated from the domain of healthcare (15 papers), though some papers did not tie their findings to any specific application. Many papers also considered users who were expert clinicians (12 papers), or non-domain experts, who were not familiar with AI (15 papers). Principles were typically created from empirical insights gathered via user studies (19 papers), or as part of literature reviews from previous work (13 papers). Finally, the majority of papers specified that whilst their design principles may originate from a specific study/domain, they could be treated as general and apply to other settings (23 papers).

The codes assigned to each principle are shown in Table 2, along with their respective meta-codes and area of focus. Figure 2 shows the count of codes across the areas of focus. Figures 3–6 show the count of principles within each area of focus and are presented within their relevant sub-sections. Each area of focus, along with its meta-codes and codes, is now discussed, with specific principles

Table 1. An Overview of the Collected 35 Papers, Including Information about the Principles Collected

Paper	Domain	Expected Users	Number of Principles	Principle Creation	Intended Application
Eardley et al. [16]	Healthcare	Non-experts	15	User study	Generic
Böckle et al. [7]	Cycling monitor app	Non-experts	3	Other's guidelines (Google)	Generic
Lee et al. [34]	Healthcare	Experts (clinical)	5	User study	Specific
Bove et al. [8]	Insurance	Non-experts	5	Authors	Generic
Long et al. [36]	Public art	Non-experts	3	User study	Specific
Oberste et al. [42]	Healthcare	Experts (clinical)	4	User study	Specific
Kim et al. [27]	Healthcare	Experts (clinical) and non-experts	7	Literature and user study	Generic
Sokol and Flach [47]	Generic	Not specified	42	Literature review	Generic
Kim et al. [28]	Weather	Experts (forecasters)	3	User study	Specific
Naiseh et al. [41]	Healthcare	Experts (clinical)	5	User study	Generic
Benjamin et al. [5]	Education	Experts (historians)	2	User study	Generic
Palladino [43]	Ethics	Non-experts	1	Literature review	Generic
Förster et al. [18]	Generic	Not specified	8	Other's guidelines (ISO)	Generic
Georgieva et al. [19]	Ethics	Not specified	9	Other's guidelines (HLEG)	Generic
Lekadir et al. [35]	Healthcare	Experts (clinical)	13	Literature review	Specific
Bunde et al. [11]	Healthcare	Experts (clinical)	1	User study	Generic
Schoonderwoerd et al. [46]	Healthcare	Experts (clinical)	19	User study	Generic
Burgess et al. [12]	Healthcare	Experts (clinical)	7	User study	Specific
He et al. [22]	Healthcare	Experts (clinical)	8	User study	Specific
Jin et al. [26]	Generic	Non-experts	17	User study	Generic
Chromik and Butz [13]	Generic	Not specified	4	Literature review	Generic
Gu et al. [20]	Healthcare	Experts (clinical)	10	User study	Specific
Bove et al. [9]	Insurance	Non-experts	3	Authors	Generic
Larasati et al. [33]	Healthcare	Non-experts	14	User study	Specific
Anderson et al. [3]	Generic	Non-experts	4	Other's guidelines ([30])	Generic
Brunotte et al. [10]	Data privacy	Non-experts	9	User study	Specific
Ahmad et al. [1]	Generic	Not specified	11	Literature and user study	Generic
Herm et al. [23]	Healthcare	Non-experts	4	User study	Generic
Wang et al. [49]	Healthcare	Experts (clinical)	15	Literature review	Generic
Herrmanny and Torkamaan [24]	Healthcare	Experts (clinical)	28	Literature review	Specific
Cirquiera et al. [14]	Insurance	Non-experts	18	Literature and user study	Specific
Alzubaidi et al. [2]	HR	Not specified	7	Literature review	Generic
Ridley [44]	Recommender systems	Non-experts	11	Literature review	Generic
Long and Magerko [37]	Education	Non-experts	8	Literature review	Generic
Schmid and Wrede [45]	Generic	Not specified	7	Literature review	Generic

quoted from papers to provide examples of the categories. When reporting the following codes, there are sometimes overlap when principles are referring to the explanations provided within an AI system, the AI system overall, or a mixture of both. For clarity, throughout this section when a code refers only to the system broadly, we use the phrasing, 'an HCXAI system should...'. When the code is specific to explanations, we use the phrasing, 'explanations provided by HCXAI systems should...'. When the code is applicable to both system and explanation, we use the phrasing, 'an HCXAI system and its explanations should...'.

4.1 The Human-Centred in HCXAI

This area of focus refers to principles that considered the human elements that influence the design of HCXAI, and consists of 54 principles (16% of all principles), as shown in Figure 3. It consists of three meta-codes: Consider Human Cognition (23 principles; 7%), Consider Human Bias (18 principles; 5%), and Understand the Context (13 principles; 4%).

Consider Human Cognition refers to a need to understand human cognitive processes in order to design an HCXAI system and its explanations that can complement them, and has four codes. The first, *User Knowledge*, has six principles, and states that explanations provided by HCXAI systems should consider and complement what knowledge and expertise the user has (e.g., 'explanations should be tailored to the specific domain expertise', Paper [27]). This may refer to their domain expertise, their knowledge of ML systems or general background knowledge. Considering what knowledge a user has makes it easier to design comprehensible explanations (e.g., 'discussing the level and type of background knowledge required to comprehend an explanation is crucial', Paper [47]). Knowledge varies between users, so it is important to consider who the intended user is

Table 2. The Codes Identified in the Content Analysis, Sorted into Meta-Codes and Areas of Focus

Area of Focus	Count	Meta-Code	Count	Code	Count
The Human-Centred in HCXAI *The human elements that influence the design of HCXAI*	54	Consider Human Cognition	23	User Knowledge	6
				Encourage Learning	6
				User Engagement	6
				Mental Model	5
		Consider Human Bias	18	Types of Bias	8
				Novelty/Abnormality	6
				Cognitive Load	4
		Understand the Context	13	Consider the User	5
				Consider the Wider Context	5
				Consider the Task	3
Design Aspects of HCXAI *The process of designing AI systems whilst keeping the user as the central focus*	51	Explain Reason for Providing Explanation	18	Is Explanation Needed	9
				Goal of Explanation	6
				Tradeoffs	3
		Consider Design Process Elements	33	Involve Users and Stakeholders	11
				Presentation	8
				Evaluation	7
				Involve Multidisciplinary Experts	5
				Iterative Design	2
The XAI in HCXAI *Details of the algorithms that generate explanations for AI systems*	137	Consider the Content of Explanation	69	Information Used in Explanation	19
				Feature Importance	12
				Case Similarity/Dissimilarity	12
				Types of Explanation	11
				Use Case Specific Information	8
				Contextual Information	7
		Consider System Justification Shown to User	32	Why System Exists	11
				General System Explanation	8
				Explain System Strengths/Weaknesses	7
				Explain Capabilities Statement	6
		Consider Explanation Justification Shown to User	21	Certainty/Confidence	8
				User Scepticism	8
				Why This Explanation	5
		Combine Explanations	15	Combine Explanations	10
				Examples of Combined Explanation	5
Characteristics of HCXAI *The elements specific to HCXAI rather than HC or XAI*	88	Characteristics of HCXAI	88	Adaptable	15
				Appropriately Simple	13
				Traceable	10
				Interactive	8
				Time Sensitive	8
				Accurate	8
				Consistent	7
				Controllable	7
				Conversational	7
				Understandable	5

(e.g., ‘the intended audience of an explainable method may vary from a domain expert, through a requirement of a general knowledge about a problem, all the way to a lay audience’, Paper [47]).

The second code, *Encourage Learning*, has six principles, and states users should learn how the HCXAI system and its explanations function via using it (e.g., ‘support training and learning’, Paper [41]). Learning can be supported by using a variety of explanation methods (e.g., ‘require a wide range of explanation to support user’s own learning’, Paper [26]) that are interactive (e.g., ‘consider including […] interactive demonstrations in order to aid in learners’ understanding of AI’, Paper [37]). Doing so can prevent cognitive overload (e.g., ‘to prevent cognitive overload, consider […] introducing scaffolding that fades as the user learns more about the system’s operations’, Paper [37]) and helps to increase trust (e.g., ‘the introduction of the AI tool is a core opportunity for trust building’, Paper [12]).

The third code, *User Engagement*, has six principles, and states an HCXAI system and its explanations should encourage the user to interact with it (e.g., ‘the user should be encouraged to […] get involved in the generation process’, Paper [24]). Motivating users to engage with explanations helps them understand how they can interact with them, which is important as it fosters human agency over decisions being made—‘the HCXAI principle of “active self-explanation” shifts the balance of power and agency toward the user. By giving the user more information and context,

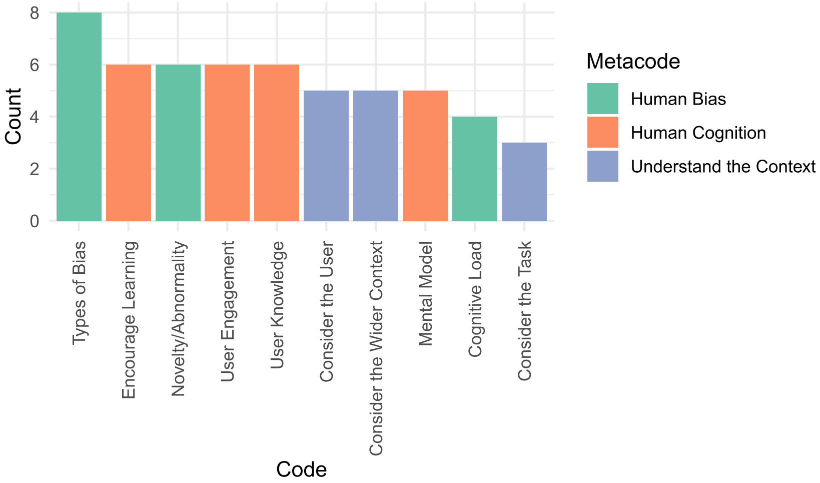

Fig. 3. Histogram of the code counts for the human-centred of HCXAI area of focus, coloured by meta-code.

they are empowered to make their own assessments and explanations rather than only receiving an algorithmic explanation', Paper [44].

The final code, *Mental Model*, has five principles, and states an HCXAI system and its explanations need to consider what information a user has and how they apply it to the current context (e.g., 'explanations should align with clinicians' cognitive processes to maximize efficiency', Paper [27]). Doing so makes it easier to know what information an explanation will need to provide (e.g., 'an explainability system has to "know" what the user knows and expects to determine the content of the explanation', Paper [47]).

Consider Human Bias refers to a need to understand the limitations in human cognition when designing explanations provided by HCXAI systems, and has three codes. The first, *Types of Bias*, has eight principles, and states the need to identify any cognitive biases that may be present when a user interacts with an explanation (e.g., 'it is important to identify any resulting bias from the use of explainability methods', Paper [35]). Examples of cognitive biases to be aware of were given as principles, such as confirmation bias (e.g., 'avoid confirmation and early closure', Paper [49]), as well as ways to avoid them (e.g., 'most explanations are not of a causal nature. If this is the case, this property needs to be explicitly communicated to the users so that they can avoid drawing incorrect conclusions', Paper [47]).

The second code, *Novelty/Abnormality*, has six principles, and states explanations provided by an HCXAI system that are different/not typical need to be clearly communicated to the user (e.g., 'explanations should contain surprising or abnormal characteristics (that have low probability of happening, e.g. a rare feature value) to point the user's attention in an interesting direction', Paper [47]). This prevents users from becoming complacent when interacting with similar explanations over time (e.g., 'design for challenging habitual actions', Paper [41]), however the explanation itself should be consistent to make novelty easier to notice (e.g., 'explanations should be familiar to the users', Paper [2]).

The final code, *Cognitive Load*, has four principles, and states explanations provided by an HCXAI system should not provide too much information at once to the user (e.g., 'do not overwhelm the user', Paper [3]). This can be done by providing short explanations (e.g., 'explanations should be [...] succinct enough to avoid overwhelming the explainee with unnecessary information', Paper [47]).

Understand the Context refers to a need to understand where the HCXAI system will be deployed, and has three codes. The first, *Consider the User*, has five principles, and states the need to understand who will use the AI system (e.g., 'the need for understanding user', Paper [18]). This includes what type of user will be involved, and what they are expected to do (e.g., 'the importance of context (regarding user objectives, decision consequences, timing, modality, and intended audience)', Paper [44]).

The second code, *Consider the Wider Context*, has five principles, and states a need to understand what affects interactions between humans, the HCXAI system, and its explanations more broadly (e.g., 'Consider organizational or project context beyond performance. Other constraints typically influence the choice of algorithm. They largely depend on environmental factors such as cost (training and inference), time constraints, end user abilities, and laws', Paper [23]). This includes considering what information an explanation should include (e.g., 'explanations should be tailored to [...] the specific contextual application of professionals', Paper [27]), and what is likely to be useful to users—'account for what is possible and realistic [...] for the clinical context', Paper [12].

The final code, *Consider the Task*, has three principles, and states a need to understand what task the HCXAI system will be involved in (e.g., 'the need for understanding [...] task', Paper [18]). This includes understanding the workflow the AI system is to be integrated into; 'Pinpoint where complex decisions need to take place in a clinical workfow versus tools that provide blanket data that physicians already know', Paper [12].

4.2 Design Aspects of HCXAI

This area of focus refers to the process of designing HCXAI by keeping the user as the central focus, and consists of 51 principles (16% of all principles), as shown in Figure 4. It consists of two meta-codes: Consider Design Process Elements (33 principles; 10%) and Explain Reason for Providing Explanation (18 principles, 5%).

Consider Design Process Elements refers to principles that outline how to develop an HCXAI system and its explanations, and has five codes. The first code, *Involve Users and Stakeholders*, has 11 principles, and states a need to include those that will interact with the tool during the design process (e.g., 'user involvement in design and development', Paper [18]). This may be the end-user, but it is important to consider other key stakeholders as well (e.g., 'comprehensively consider multiple stakeholders with different interests in the AI system', Paper [22]), and if these differing perspectives conflict—'measure whether a design solution that meets the explanation needs of one stakeholder will harm the interests of other stakeholders', Paper [22].

The second code, *Presentation*, has eight principles, and refers to the layout and interface design of the explanations provided by an HCXAI system (e.g., 'consider mode(s) of explanation presentation', Paper [16]). Features that may affect processing of information (e.g., 'design for accessibility', Paper [16]) as well as how information is communicated, are considered here—'design for inclusiveness [...] it is important to consider measurement units, language, images, and visuals', Paper [16]. The language used by HCXAI should also be considered (e.g., 'consider complementing implicit explanations with rationales in natural language', Paper [13]).

The third code, *Evaluation*, has seven principles, and states a need to test an HCXAI system design with users (e.g., 'rigorous empirical evaluation', Paper [45]). Doing so helps to increase the robustness and safety of the system (e.g., 'performance in various testing cases to show AI's robustness on safety', Paper [26]).

The fourth code, *Involve Multidisciplinary Experts*, has five principles, and states a need to include experts from a variety of fields during development (e.g., 'a multidisciplinary team with diverse skills and perspectives is required', Paper [18]). This includes those that helped to build the system (e.g., 'involve the technical expert team members in review of the design development', Paper [16]),

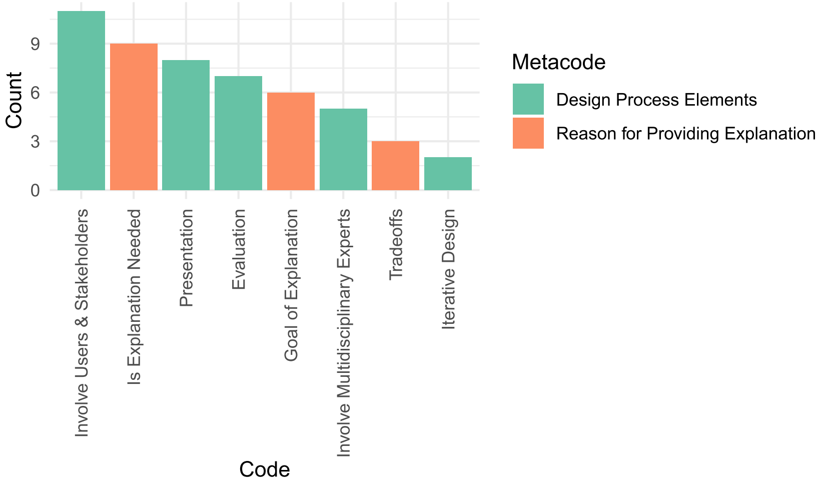

Fig. 4. Histogram of the code counts for the design aspects of HCXAI area of focus, coloured by meta-code.

but their input must be considered alongside user feedback—'balance the feedback given by the users and experts', Paper [16].

The final code, *Iterative Design*, has two principles, and states a need to perform multiple design stages to ensure the HCXAI system is designed correctly for the context (e.g., 'the need for an iterative process', Paper [18]).

Explain Reason for Providing Explanation states the importance of explicitly stating why an HCXAI system is being built, and has three codes. The first, *Is Explanation Needed*, has nine principles, and states a need to identify if an HCXAI system requires an explanation for a specific task (e.g., 'does an explainee need to understand the inner workings of a predictive model?', Paper [47]). Depending on the context an explanation may not in fact be needed (e.g., 'explanations are not always necessary', Paper [44]), and the extent of explanation needed is also likely to be context-dependent—'consider the degree of explanation that end users need', Paper [23].

The second code, *Goal of Explanation*, has six principles, and refers to the need to identify what the explanation provided by the HCXAI system is trying to achieve (e.g., 'explanation content depends on specific explanation goals', Paper [26]). This involves understanding why explanations exist within human communication (e.g., 'the need for an explanation is not about "why" or "how" but rather for a set of confidence or performance metrics', Paper [44]).

The final code, *Tradeoffs*, has three principles, and states a need to outline what is feasible for an explanation provided by an HCXAI system to deliver given the context (e.g., 'trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)', Paper [19]). This may also involve considering user cognition to understand what should and should not be traded off (e.g., 'balancing the tradeoff between coherence with the explainee's mental model, novelty and overall plausibility', Paper [47]).

4.3 The XAI in HCXAI

This area of focus refers to the algorithms generating the explanation for AI systems, and consists of 137 principles (42% of all principles), as shown in Figure 5. It consists of four meta-codes: Consider the Content of Explanation (69 principles; 21%), Consider System Justification Shown to User (32 principles; 10%), Consider Explanation Justification Shown to User (21 principles; 6%), and Combine Explanations (15 principles; 5%).

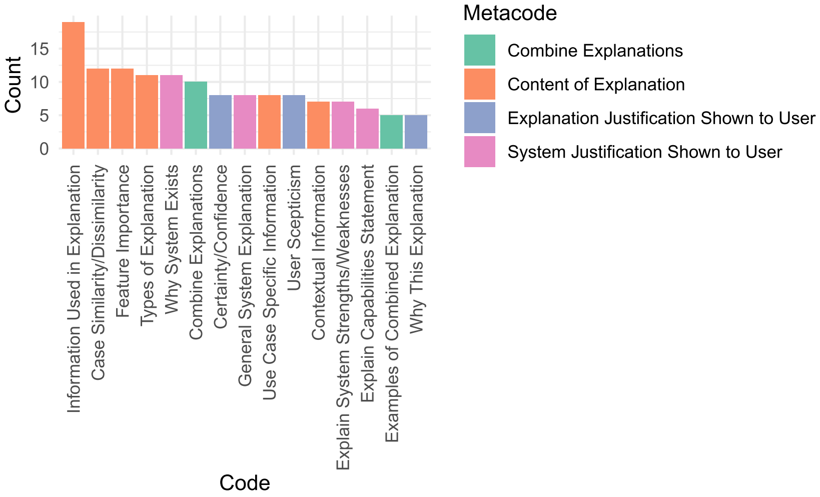

Fig. 5. Histogram of the code counts for the XAI of HCXAI area of focus, coloured by meta-code.

Consider the Content of Explanation refers to what information an HCXAI system needs access to and what explanations it should show to users, and has six codes. The first, *Information Used in Explanation*, has 19 principles, and describes what types of information within the explanations provided by an HCXAI system should include (e.g., 'the information that is used to make the classification', Paper [46]). This includes reference information (e.g., 'providing any kind of reference or value for orientation', Paper [24]), the inputs of the explanation (e.g., 'the data user inputted to the system', Paper [33]) and the output (e.g., 'system result, e.g., pre-diagnosis, analysis, recommendation', Paper [33]).

The second code, *Feature Importance*, has 12 principles, and refers specifically to explanations provided by HCXAI systems using feature importance techniques (e.g., 'the XAI model should easily explain which features or variables influenced the model's predictions', Paper [2]). The relationships between features should also be explained (e.g., 'enables experts to observe [...] the relationships between features in the dataset', Paper [14]), as well as how these features would change between classifications—'from what value of feature X the classification would have been different,' Paper [46].

The third code, *Case Similarity/Dissimilarity*, has 12 principles, and refers specifically to explanations provided by HCXAI systems using case similarity (e.g., 'how this case relates to a specific similar case', Paper [46]) and dissimilarity (Paper [14]). This could involve highlighting the similarities between cases (Paper [49]) or showing the range of cases that are included in the same classification (e.g., 'the most different cases with the same classification', Paper [46]).

The fourth code, *Types of Explanation*, has 11 principles, and describes the types of explanations provided by HCXAI systems that can currently be provided (e.g., 'the power of contrastive examples and approaches', Paper [44]). Examples given include the use of counterfactual explanations (e.g., 'facilitate sensitivity analysis with What If explanations to test stability of primary hypothesis', Paper [49]), 'what if' explanations (e.g., 'facilitate sensitivity analysis with What If explanations to test stability of primary hypothesis', Paper [49]), and graphics—'consider including graphical visualizations', Paper [37].

The fifth code, *Use Case Specific Information*, has eight principles, and describes information that is specific to the use case of the study it was collected from (e.g., 'general disease information, e.g., name, symptoms, caused', Paper [33]). Many of these are specific to clinical contexts (e.g., 'compare disease with prototypes of the condition', Paper [49]). These are likely only relevant to the context being studied, and so are unlikely to generalise to all explanations provided by HCXAI systems.

The final code, *Contextual Information*, has seven principles, and refers to the need for explanations provided by HCXAI systems to include information that is relevant to the context (e.g., 'provide information related to the context of the current operation', Paper [22]). This also includes considering how to pair explanations provided by the XAI system with the current context (e.g., 'supplement explanation methods with contextual cues', Paper [5]).

Consider System Justification Shown to User refers to a user's desire to query the HCXAI system's purpose, and has four codes. The first, *Why System Exists*, has 11 principles, and refers to a need to explain to users the HCXAI system's purpose (e.g., 'the rationale for deploying it, should be available', Paper [19]). This includes how the design was chosen (e.g., 'design choices of the system [...] should be available', Paper [19]), and where it is expected to be used ('every explainability approach should be accompanied by a list of its intended applications', Paper [47]), to help users understand its purpose—'explain why the platform is useful to your end users', Paper [16].

The second code, *General System Explanation*, has eight principles, and refers to a need to explain what the HCXAI system does overall (e.g., 'general system information, e.g., data, system accuracy', Paper [33]). This includes how user data are used (e.g., 'explain to the user how their information is used', Paper [1]).

The third code, *Explain System Strengths/Weaknesses*, has seven principles, and refers to a need to explain what the HCXAI system performs well at (e.g., 'strengths of the model must be explained', Paper [2]) and its limitations (e.g., 'explain limitations of the AI', Paper [1]). This can be done by explaining what the model output is based on (e.g., 'the user will better understand the limitations of an explanation if it is accompanied by all the necessary conditions for it to hold', Paper [47]).

The final code, *Explain Capabilities Statement*, has six principles, and refers to a need to explain what the HCXAI system is capable of achieving (e.g., 'explain system functionalities', Paper [1]). This is similar to *Explain System Strengths/Weaknesses*, but refers specifically to explaining what the AI system should be used for given its design. It includes explaining the underlying algorithm (e.g., 'system algorithm or the technical process to gets its results', Paper [33]) in order to increase transparency—'give transparency on the ML system's scope and basic operations to provide guidance on how to interpret explanations', Paper [8].

Consider Explanation Justification Shown to User refers to a user's desire to question the specific explanation provided by the HCXAI system for its decision, and has three codes. The first code, *Certainty/Confidence*, has eight principles, and refers to a need for the system to provide a metric for how certain it is of its recommendation (e.g., 'indicate decision certainty level', Paper [26]). This can be achieved via many metrics (e.g., 'indicating the extent of the system's uncertainty through parameters, such as error, accuracy, precision, confidence etc.', Paper [24]).

The second code, *User Scepticism*, has eight principles, and refers to enabling the user to question and interrogate an explanation provided by the HCXAI system (e.g., 'the user should be encouraged to question [...] the system result', Paper [24]). This can be done to check the correctness of the result (e.g., 'for the user to check the input (is it correct or not)', Paper [33]) or because the user is unsure of the result—'need explanations for verification', Paper [26].

The final code, *Why This Explanation*, has five principles, and refers to a need to explain how a specific explanation provided by the HCXAI system was achieved (e.g., 'provide local explainability', Paper [20]). This may include why an explanation was given over another (e.g., 'why it is this classification, and not another one', Paper [46]), or why it is a better explanation—'Provide insight into why a suggestion [...] is superior to others (in contrast to how it was generated)', Paper [24].

Combine Explanations refers to a need for HCXAI systems to provide more than one type of explanation, and has two codes. The first, *Combine Explanations*, has 10 principles, such as 'consider offering multiple explanation methods and modalities to enable explainees to triangulate insights', Paper [13]. Alongside this, it is important to explain how the explanations are connected (e.g., 'present the logical relationship that connects these multiple criteria/features/sources of information', Paper [20]). The second, *Examples of Combined Explanation*, has five principles, and describes specific examples of which explanations an HCXAI system can combine (e.g., 'show input attributions for multiple outcomes to allow contrastive reasoning', Paper [49]), and what reference information should be included—'pair each local feature importance explanation with global information provided by a domain expert. It should provide some brief justification about how a feature might impact the prediction regardless of its value', Paper [8].

4.4 Characteristics of HCXAI

This area of focus refers to characteristics that are specific to HCXAI rather than just XAI or human elements, and consists of 88 principles (27% of all principles), as shown in Figure 6. They are achieved by considering all aspects of HCXAI in combination. It contains 10 codes all under a single meta-code.

The first code, *Adaptable*, has 15 principles, and states an HCXAI system and its explanations should be flexible to the user (e.g., 'make an adaptive system for [...] collaborative decision making', Paper [34]). This includes adapting to a specific user's needs (e.g., 'users should be able to customise the explanation that they get to suit their needs', Paper [47]), which may vary between users (e.g., 'explanations depend on different stakeholders', Paper [26]). Further, adapting the complexity of an explanation to a user's preference is important—'the complexity of explanations should be tuned to the recipient', Paper [47].

The second code, *Appropriately Simple*, has 13 principles, and states explanations provided by an HCXAI system should be at the correct level of complexity for the user and the context (e.g., 'if the system does not allow for explanation complexity to be adjusted by the user, it should be as simple as possible by default (unless the explainee explicitly asks for a more complex one)', Paper [47]). Simple may refer to the language used (e.g., 'simple and General Uncomplicated wording that is acceptable for laypeople from various education background and level', Paper [33]), the length/detail of the explanation (e.g., 'avoid overly detailed explanations', Paper [22]), or the knowledge users are expected to have to understand the explanation—'explanations should be [...] simple in terms of audience knowledge', Paper [2].

The third code, *Traceable*, has 10 principles, and states an HCXAI system and its explanations should show how an explanation was generated (e.g., 'if possible, every explanation should be accompanied by an explainability trace indicating which training data points were influential for a prediction and the role that the model and its parameters played', Paper [47]). This allows users to understand how a decision was reached in terms of evidence (e.g., 'there should be explainability [...] locally (what evidence leads to the computed result of each criterion)', Paper [20]) as well as the AI's decision-making process—'inference path to provide the reasoning process of the AI model', Paper [14].

The fourth code, *Interactive*, has eight principles, and states an HCXAI system and its explanations should not be static but rather interactable by the user (e.g., 'interaction for [...] human-in-the-loop decision making', Paper [45]). This can take the form of providing feedback to the user (e.g., 'providing system feedback to the user also supports fluent interaction', Paper [24]), as well as providing further information when needed (e.g., 'the user to request detailed information', Paper [33]).

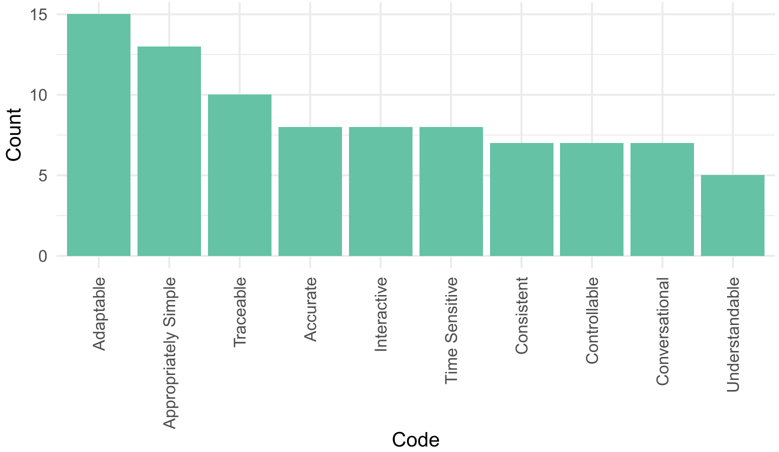

Fig. 6. Histogram of the code counts for the characteristics of HCXAI area of focus.

The fifth code, *Accurate*, has eight principles, and states explanations provided by an HCXAI system should present accurate information given the context (e.g., 'the content of the explanation needs to be accurate', Paper [16]). This also considers how complete an explanation is (e.g., 'be complete', Paper [3]), and the accuracy of the interpretation of the AI predictive model—'how truthful an explanation is with respect to the underlying predictive model', Paper [47].

The sixth code, *Time Sensitive*, has eight principles, and states an HCXAI system should consider when its explanations are given (e.g., 'explanation should be timely', Paper [19]). This includes: where in a user's interaction/decision-making process the explanation should be presented (e.g., 'the XAI model should easily explain [...] at what step the decision is made', Paper [2]), the up-to-dateness of the underlying data (e.g., 'the content of the explanation [...] needs to be up-to-date', Paper [16]), and the need to highlight any changes—'any updates and changes to the platform need to be included and made available', Paper [16].

The seventh code, *Controllable*, has seven principles, and states an HCXAI system should be controlled by the user and not undermine their agency (e.g., 'give the clinician agency/control over model output', Paper [12]). The user should be able to control the explanation process (e.g., 'the user should be enabled [...] to exert influence on the system', Paper [24]) to allow for effective oversight—'to ensure efficient oversight and decision-making, humans should maintain control over AI systems', Paper [27].

The eighth code, *Consistent*, has seven principles, and states an HCXAI system and its explanations should be presented in a recognisable and consistent way (e.g., 'consistency of explanations', Paper [45]). This relates to the reliability of outputs (e.g., 'users are interested in the local reliability of the predictions', Paper [28]) and the performance (e.g., 'users require stated and observed performance', Paper [26]) in a variety of applications—'require AI to maintain the same capability and perform well for minority subgroups', Paper [26].

The ninth code, *Conversational*, has seven principles, and states explanations provided by the HCXAI system should take the form of a dialogue between the system and the user following the conventions of human communication (e.g., 'the explanation process should [...] be "social" (bidirectional communication is preferred to one-way information offloading)', Paper [47]). This also includes the ability for users to correct previous mistakes (e.g., 'corrigibility for human-in-the-loop decision making', Paper [45]) and to ask for further clarification—'consider offering hierarchical or iterative functionalities that allow followups on initial explanations', Paper [13].

Table 3. An Overview of What Codes Were Formed from Which Domains Based on the 35 Papers Reviewed

	Generic (7 papers)	Ethics (2 papers)	Privacy (1 paper)	Health (15 papers)	Device				Business			
Code					Smarthome Health (1 paper)	Phone app Cycling (1 paper)	Public Art (1 paper)	Recommender Systems (1 paper)	HR (1 paper)	Insurance (2 papers)	Education (2 papers)	Weather (1 paper)
Accurate	✓		✓		✓							
Adaptable	✓	✓		✓						✓		
Appropriate Simplicity	✓		✓	✓	✓	✓			✓	✓		
Case Similarity/Dissimilarity	✓			✓						✓		
Certainty/Confidence	✓			✓		✓				✓		
Cognitive Load	✓										✓	
Combine Explanations	✓			✓				✓			✓	
Consider the Task	✓			✓								
Consider the User	✓			✓	✓			✓				✓
Consider the Wider Context	✓			✓							✓	
Consistent	✓			✓								✓
Controllable	✓			✓								
Conversational	✓			✓								
Design Evaluation	✓			✓								
Encourage Learning	✓			✓							✓	
Examples of Combined Explanation				✓						✓		
Explain Capabilities Statement	✓			✓						✓		
Explain System Strengths/Weaknesses	✓			✓					✓	✓		
Feature Importance	✓			✓					✓	✓		
General System Explanation	✓		✓	✓						✓		
Goal of Explanation	✓							✓				
Human Bias	✓			✓								
Information Used in Explanation	✓			✓				✓		✓		
Interactable	✓			✓			✓			✓		
Involve Multidisciplinary Experts	✓			✓	✓							
Involve users and stakeholders	✓			✓				✓				
Is Explanation Needed	✓			✓				✓				
Iterative Design	✓											
Mental Model	✓			✓		✓						
Novelty/Abnormality	✓			✓					✓			
Presentation	✓		✓	✓	✓							
Time Sensitive	✓	✓		✓	✓				✓			
Traceable	✓	✓		✓			✓			✓	✓	
Tradeoffs	✓	✓										
Types of Explanation	✓			✓				✓		✓	✓	
Understandable		✓		✓			✓					
Use Case Specific Information				✓								
User Engagement				✓				✓				
User Knowledge	✓			✓								
User Scepticism	✓			✓								
Why System Exists	✓	✓	✓	✓	✓							
Why This Explanation	✓			✓								

The final code, *Understandable*, has five principles, and states an HCXAI system and its explanations should be understandable to users who may not have in-depth knowledge of how ML systems work (e.g., 'decisions made by an AI system can be understood [...] by human beings', Paper [19]). This includes how interface design may influence understandability (e.g., 'considering how interface design plays a role in understandability', Paper [36]) as well as how users will understand what to do when given an explanation—'the user should be enabled to understand [...] its implications', Paper [24].

4.5 Influence of Principle Specificity, Domain and Intended Users

Given that the above principles were formed from a varied set of papers in terms of their intended domain and user, this section explores how the identified codes relate to these factors. Most papers provided generic principles rather than specific to an intended application (23/35 papers). However, all codes were formed from a combination of both generic and specific principles, except in nine cases: Cognitive Load, Goal of Explanation, Is Explanation Needed, Iterative Design, Mental Model, Novelty/Abnormality, Tradeoffs, Types of Explanation, and User Knowledge. In these cases, only principles intended to be generic formed these codes. Therefore, whilst specific principles were used in the dataset, none of them formed their own unique categories.

In terms of the domains studied, healthcare was the most common (15 out of 35 papers), followed by no specified domain, here considered to be generic domain application (7/35 papers). An overview of how the generated codes relate to the intended application domain is shown in Table 3.

Principles not specific to a domain (i.e., generic application) aided in generating all codes, apart from four: Examples of Combined Explanation, Use Case Specific Information, Understandable, and User Engagement. In terms of the former two this is understandable, as these were codes identified specific information about the study within the paper, and so they would not be expected to appear here. In terms of the latter two, Understandable applies to the domains of ethics, health, and public

Table 4. An Overview of What Codes Were Formed from Which Intended Users Based on the 35 Papers Reviewed

		Experts		Mixed	Non-Domain Experts (14 Papers)	Generic (7 Papers)
Code	Forecasters (1 Paper)	Clinical (11 Papers)	Historians (1 Paper)	Experts (Clinical) and Non-Domain Experts (1 Paper)		
Accurate					✓	✓
Adaptable		✓		✓	✓	✓
Appropriate Simplicity		✓			✓	✓
Case Similarity/Dissimilarity		✓			✓	✓
Certainty/Confidence		✓			✓	
Cognitive Load					✓	✓
Combine Explanations		✓	✓		✓	✓
Consider the Task		✓				✓
Consider the User	✓				✓	✓
Consider the Wider Context		✓	✓	✓	✓	✓
Consistent	✓	✓			✓	✓
Controllable		✓		✓	✓	✓
Conversational		✓			✓	✓
Design Evaluation		✓			✓	✓
Encourage Learning		✓			✓	
Examples of Combined Explanation		✓			✓	
Explain Capabilities Statement		✓		✓	✓	✓
Explain System Strengths/Weaknesses				✓	✓	✓
Feature Importance		✓			✓	✓
General System Explanation		✓			✓	✓
Goal of Explanation					✓	
Human Bias		✓				✓
Information Used in Explanation		✓			✓	✓
Interactable		✓			✓	✓
Involve Multidisciplinary Experts		✓			✓	✓
Involve Users and Stakeholders		✓			✓	✓
Is Explanation Needed					✓	✓
Iterative Design					✓	✓
Mental Model				✓	✓	✓
Novelty/Abnormality		✓				✓
Presentation		✓			✓	✓
Time Sensitive					✓	✓
Traceable		✓			✓	✓
Tradeoffs						✓
Types of Explanation		✓			✓	✓
Understandable		✓			✓	✓
Use Case Specific Information		✓			✓	
User Engagement		✓			✓	
User Knowledge				✓		✓
User Scepticism		✓			✓	
Why System Exists		✓			✓	✓
Why This Explanation		✓			✓	

art, and User Engagement applies to the domains of health and recommender systems. Therefore, the generated codes from this analysis apply across several domains, though an HCXAI being Understandable and considering User Engagement have more specific applicability.

In terms of the intended users studied, many applied to both clinical experts (12/35) or non-domain experts (15/35), whilst many papers also provided no information on the intended user in terms of expertise (7/35). Here, this unspecified user type is considered to provide generic principles, as principles created from these papers should apply regardless of the level of expertise of the user. An overview of how the generated codes relate to the intended users is shown in Table 4.

By combining the three types of Expert User together (generic, clinical, and historians), only 10 codes did not form from this category: Accurate, Cognitive Load, Explain System Strengths/Weaknesses, Goal of Explanation, Is Explanation Needed, Iterative Design, Mental Model, Time Sensitive, Tradeoffs, and User Knowledge. This is different to the Non-Domain Expert User category, which did not form five codes: Consider the Task, Human Bias, Novelty/Abnormality, Tradeoffs, and User Knowledge. This indicates high overlap between codes relevant to both Experts and Non-Domain

Experts, as only Tradeoffs and User Knowledge did not form from either category. In contrast, eight codes were not formed from the Generic User category: Certainty/Confidence, Encourage Learning, Examples of Combined Explanation, Goal of Explanation, Use Case Specific Information, User Engagement, User Scepticism, and Why This Explanation. Only Goal of Explanation overlaps with Expert Users here, indicating there are only a few codes specific to a specific user group (7/42 codes). Therefore, the codes generated from this analysis can be readily applied to most user types.

5 Discussion

The aim of this study was to analyse what HCXAI design principles have been suggested to date, to distil the literature into a summary set of design principles. Alongside this, what the field currently means by the term 'HCXAI' is explored. This was done by performing a scoping review of the extracted design principles. To this end, two research questions were answered.

5.1 RQ1: What HCXAI Design Principles Currently Exist and What Methods Are Used to Identify Them?

A Content Analysis revealed 43 types of design principle, sorted into 10 groups in 4 areas of focus. The areas of focus reflected the different components within HCXAI design principles: the human-centred element, the design element, the XAI element, and the characteristics of HCXAI which are enabled by combining the previous elements. The characteristics of HCXAI were found to be an AI system that is adaptable, appropriately simple, traceable, interactive, time sensitive, accurate, consistent, controllable, conversational, and understandable.

The high volume of codes suggests HCXAI design principles are highly varied and cover a complex range of concepts. Consequently, they could not be easily summarised without the use of meta-codes. By extension this suggests HCXAI is a complex concept to achieve, requiring multiple areas of focus that need to be combined and work together. This complexity and its implications for the field are further explored in the following sub-section.

In terms of the methods used, the majority of papers used user studies to elicit design principles (19/35). Papers also commonly conducted literature reviews to collate previous design principles (13/35). In rare cases, papers would combine both a user study and literature review (3/35), make use of previous guidelines from other fields or papers (4/35), or be suggested by the authors themselves (2/35). Overall, this suggests the creation of HCXAI design principles is still in the infant stages, where new principles are derived from empirical studies rather than from an established and agreed upon set. Implications for the field based on its infancy are explored in the following sub-section.

Overall, as explored in Section 4.5, the level of specificity, domain and intended user had little effect on code generation, as most were formed evenly from all of these variables. This indicates, *whilst the papers sampled considered HCXAI in diverse ways, the codes generated from the analysis reflect general HCXAI principles that apply across domains and user types.* Given this, a summary of the HCXAI principles currently available in the field are shown in Table 5, based on each code generated from the analysis.

5.2 RQ2: What Insights about the Field of HCXAI Can Be Gained from These Design Principles?

The second research question revealed three key insights: (1) a need for clarity on what design principles are specific to HCXAI, as opposed to XAI, human factors, or user-centred design; (2) a need for clarity over the definition of HCXAI itself; and (3) the identification of 10 key characteristics of HCXAI.

Table 5. Design Principles for HCXAI Summarised from the Content Analysis

Code	Summary Principles—An HCXAI should...
Accurate	Present a complete and accurate explanation of the AI predictive model
Adaptable	Be flexible to the user, their specific needs, and their preferences for complexity
Appropriate Simplicity	Be the correct level of complexity for the user and context in terms of language used, explanation length/detail, and expected user knowledge
Consider the Wider Context	During development, understand what factors affect human–AI system interactions more broadly, to inform what information to include
Case Similarity/Dissimilarity	Highlight the similarities between cases or show the range of cases included in the same classification
Certainty/Confidence	Provide a metric for how certain it is of a provided explanation being correct
Cognitive Load	Not provide too much information at once to the user, for example by providing short explanations
Combine Explanations	Provide more than one type of explanation and explain how the explanations are connected
Consider the Task	During development, understand what task the AI system is to be involved in, such as the workflow it will be integrated into
Consider the User	During development, understand what types of users will use the AI system, and what they are expected to do with it
Consistent	Present explanations in a recognisable and consistent way, so that the output/performance is reliable
Contextual Information	Provide information that is relevant to the context
Controllable	Be controlled by the user and not undermine their agency, to allow for effective oversight
Conversational	Take the form of a dialogue between the AI system and user, following the conventions of human communication
Design Evaluation	During development, test the design with users to increase the robustness and safety of the system
Encourage Learning	Help users learn how the system works using various interactive explanation methods to prevent cognitive overload and help increase trust
Examples of Combined Explanation	[Descriptive code rather than a principle]
Explain Capabilities Statement	Explain what the system is capable of achieving, what the AI system should be used for, and its underlying algorithm to increase transparency
Explain System Strengths/Weaknesses	Explain what the AI system performs well at and what are its limitations
Feature Importance	Use feature importance techniques, explain relationships between features, and how features would change between classifications
General System Explanation	Explain what the system does overall, including how user data are used
Goal of Explanation	Identify what the explanation of the AI system is trying to achieve, using human communication as a starting point
Human Bias	During development, identify the types of cognitive bias that may be present and ways to avoid them
Information Used in Explanation	During development, consider the types of information to show to users, and what to include alongside this information
Interactable	Be interactable by the user, such as by providing feedback or further information to the user when needed
Involve Multidisciplinary Experts	During development, include experts from a variety of fields, such as those that helped build the system
Involve Users and Stakeholders	During development, include the end user and other key stakeholders, and identify any conflicts between their perspectives
Is Explanation Needed	During development, identify if an AI system is required for a specific task/context
Iterative Design	During development, perform multiple design stages to ensure it is designed correctly for the context
Mental Model	During development, consider what information a user has and how they apply it to the current context
Novelty/Abnormality	Clearly communicate to the user results that are different/not typical via consistent explanations that make it easy to highlight novelty
Presentation	During development, consider the layout and interface design and how these may affect information processing
Time Sensitive	Be aware of when explanations are given in a user's interaction/decision-making process, as well as the up-to-dateness of data
Traceable	Show how an explanation was generated, so users can understand how a decision was reached
Tradeoffs	During development, outline what is feasible for an AI system to deliver given the context
Types of Explanation	[Descriptive code rather than a principle]
Understandable	Be understandable to users who may not have in-depth knowledge of how ML systems work
Use Case Specific Information	[Descriptive code rather than a principle]
User Engagement	Engage the user to encourage interacting with it, to help foster human agency over decisions being made
User Knowledge	Consider and complement what knowledge and expertise the user has and how this varies between users
User Scepticism	Enable the user to question and interrogate a given explanation to check the correctness of the result or investigate any unsurety they have
Why System Exists	Explain to users the AI's purpose, including how the design was chosen and where it is expected to be used
Why This Explanation	Explain how a specific explanation was achieved, such as why an explanation was given over another

5.2.1 A Need for Clarity on What Design Principles Are Specific to HCXAI. Firstly, despite care being taken to only include HCXAI design principles during paper screening, the analysis revealed areas of focus that reflected each component of HCXAI separately (i.e., the human-centred focus and XAI focus), alongside the characteristics specific to HCXAI. This suggests a high overlap between design principles created to address the concepts of human-centred design, explainable AI, and HCXAI, which blurs the boundaries of where HCXAI begins and its supporting fields end.

Furthermore, roughly a quarter of principles were specific to the characteristics of HCXAI (26.7%), which represent the core of the design principles. This means most principles instead discuss surrounding factors that affect HCXAI, such as the human-centred and XAI components, further blurring the boundaries of HCXAI. It is possible the term 'HCXAI' is temporary in nature as the field continues to mature the wider concepts around explainable AI. At this point, it may become a redundant term, as it can be better understood from a different perspective. For example, *taking a systems approach would involve considering the wider context in which an AI system exists.* Doing so may help explain how to design interactions that occur between the different human and AI elements in the system. Until then, future research may be required to clarify what is specific about HCXAI, that is distinct from human-centred or XAI design principles.

However, the existence of areas of focus within HCXAI design principles also reveals where current research has put the most emphasis. There has been a large focus on the XAI component of HCXAI, as this area contained the highest number of principles. The smallest area was the design component, followed by the human-centred component. This suggests whilst HCXAI principles may have been intended to combine human-centred design concepts with XAI, there has been a smaller research focus on how to design for this. This may affect the usefulness of HCXAI design principles currently suggested by the field. Further, it indicates more research is required into the design and human-centred components of HCXAI to further improve suggested design principles. As discussed above, taking a systems approach to the design of AI systems means it is important to consider all aspects of the context in which the AI system will exist. Doing so will lead to more appropriate and well-integrated AI systems.

5.2.2 A Need for Clarity over the Definition of HCXAI Itself. Secondly, and building on the first insight, there is some confusion over what the term HCXAI itself refers to, as seen by the varied types of codes generated. For example, some principles consider how to *design* an AI system, whilst others capture the *interactions* that happen between the human and the AI system. Other principles are concerned with the *presentation* of explanations, or identifying which XAI *techniques* are most appropriate. Despite the high variety of concepts between these codes, it was common to see multiple types of principle within the same paper. For example, Paper [49] discusses principles that reflect both the Considering Human Cognition and Combine Explanations meta-codes, which fall into the Human-Centred in HCXAI and XAI in HCXAI areas of focus respectively. Indeed, only 3 papers discussed 1 meta-code [11, 36, 43], and only 10 were exclusively related to 2 meta-codes. On average, each paper discussed principles relating to 4 of the 10 meta-codes, and 3 of the 4 areas of focus.

Given this, it can be seen that papers currently available in the field consider HCXAI from numerous overlapping angles. Whilst the varied codes seen are related concepts for HCXAI, it implies there are multiple reasons for *why* HCXAI design principles are created, and for what purpose. It is consequently important to synthesise these different ideas into one combined definition, to accurately capture the current focus of the field. Therefore, the definition of HCXAI is revisited here. At the beginning of this article, the working definition of HCXAI was described as taking into account factors that influence human decision-making and communication during the design and development process of AI systems. Given the findings of the scoping review, a second insight of this work is a reframing of the HCXAI definition:

> HCXAI refers to '*a design process to support users' understanding of an AI system by enabling meaningful interactions with the AI that are appropriate for the given context. This is supported by, but not limited to, XAI techniques and participatory engagement with key stakeholders throughout the design process. Explanations are consequently not considered a static output, but rather an emergent property of these interactions. As a result, the explanations within an AI system should exhibit one or more of the following key design characteristics: be adaptable, appropriately simple, traceable, interactive, time sensitive, accurate, consistent, controllable, conversational, and understandable.*'

Whilst it may not be possible—or desirable—for all AI systems to demonstrate all 10 of these central variables in all contexts, this does not mean such AI systems would not be considered human-centric in nature. Rather, considering these central variables when designing an AI system increases the chances that the interactions with the AI will be appropriate for the given context, which in turn will support user understanding.

Overall, this definition highlights that HCXAI is not another type of XAI technique with a focus on user-centredness, but rather a design process. It therefore considers wider concepts alongside what is being explained, such as how the AI system should be designed and what human elements require consideration during development. In turn, this implies new XAI techniques may sometimes be required to achieve the characteristics of HCXAI, whilst other times existing XAI techniques are sufficient. Knowing when a new technique may be needed requires an understanding of the context and aims of the AI system overall, as whilst some XAI techniques are naturally more human-centred than others, this is also dependent on the context [23]. There is no current consensus that one XAI technique is the most human-centred in all contexts and all applications, and so there can be no one size fits all approach to designing HCXAI. Therefore, this definition of HCXAI highlights both the context of the AI system and the chosen XAI technique must be considered simultaneously.

5.2.3 The Identification of 10 Key Characteristics of HCXAI. Finally, despite the confusion discussed above around what is specific to HCXAI, this study revealed it is still possible to identify key characteristics of an AI system built using an HCXAI design approach. *The 'Characteristics of HCXAI' area of focus can be considered the core HCXAI design principles the field currently considers*, which are supported and further nuanced by the other areas of focus in terms of context and ways to achieve these characteristics. In doing so, the entire concept of HCXAI is brought together. For example, in order for a system to be Appropriately Simple, it is important to understand what is meant by appropriate and simple, which relates to the Human aspect of HCXAI (such as the Cognitive Load code). It is possible to understand this Human aspect by focusing on how to Design the AI system with a user-centred approach, such as via the Is Explanation Needed code. Finally, for a system to be Appropriately Simple, an AI system is required, which can be designed via the Types of Explanation code. Therefore, the core characteristics of HCXAI are supported and achieved by each of the three individual components of HCXAI.

5.3 Limitations and Future Work

There are a number of limitations to note for this study. Firstly, whilst a high volume of papers were included in the initial search, there is always the possibility that papers and their principles were missed. They may not have been indexed with the ACM Digital Library or Google Scholar, and so could not have been included. Further, it is not possible to include all related search terms. For example, 'interpretable machine learning' may have provided more relevant results for consideration. Other examples include different terminology for design principles/requirements, such as 'guidelines' and 'implications for design'. Future research can use the findings of this study as a benchmark of what is currently known, so that new principles are more likely to add to existing knowledge rather than overlap with already identified principles.

Secondly, due to the nature of qualitative analysis, the categories and codes described here are based on researcher interpretation. Whilst an independent second coder was included to improve reliability of the codes, their lack of experience in XAI is a limitation. Further, other researchers assessing the same principles could, and likely would, find different categories to describe the data. However, whilst this means there can never be an 'objective' summary of HCXAI design principles, the benefits of performing a Content Analysis are its transparency and traceability. By providing the raw data (i.e., the principles) and the categories created from them in the Supplementary Materials, it is possible to observe exactly how the analysis took place, especially in terms of how each individual principle feeds into the final HCXAI design principle categories. This is in contrast to typical meta-reviews conducted in the field, where the summarising of literature is less well defined, meaning readers must take the results at face value. We encourage other researchers to analyse the principles found here to reveal new insights, and to inspire more nuanced discussions around the topic of HCXAI.

Thirdly, it is important to note the difference between describing/summarising the current state of the HCXAI field, vs. identifying what is required of HCXAI design principles moving forward. By analysing design principles in research papers it was possible to achieve the former, however this analysis alone cannot answer the latter. Whilst some areas which have received less attention can be identified, such as a noted lack of focus on the human component of HCXAI, other areas that could further improve HCXAI design principles cannot be known. Future research could therefore use the design principles identified here as a starting point for inspiring new areas of focus.

There are a number of further future research directions that can be taken from the current work. Firstly, whilst 10 key characteristics of HCXAI have been identified here, there is a need for more guidance on which characteristics are needed and when, given the specific context of the AI system being developed. For example, tradeoffs are likely between characteristics depending on the context, such as in time-critical systems where providing explanations that are time sensitive may be more important than explanations that are traceable. Secondly, future guidance is needed on how to implement these characteristics, alongside ways to evaluate/operationalise the characteristics for a given context. For example, whilst it is important that an AI system is appropriately simple, how to design an explanation that is appropriately simple, and how to reliably evaluate this simplicity, is harder to define. New methods for assessing the characteristics of HCXAI consequently may be required, alongside methodologies for how to integrate and assess the characteristics of HCXAI during the development process. This may also require the use of case studies to showcase the utility in having such a methodology for implementing HCXAI.

6 Conclusions

The field of HCXAI has produced a wide variety of design principles for developers and researchers to consider when building AI systems. By conducting a scoping review followed by a content analysis, it was possible to transparently analyse and summarise 330 principles currently suggested by the field. Ten characteristics of HCXAI were identified, supported by 43 codes from each individual element that consitutes the concept of HCXAI. Future research can use the present study as a benchmark of what the field currently considers to be the main design principles of HCXAI, as well as identify new areas that are needed to improve these principles further.

Acknowledgement

We thank Lauren Winter for providing the independent secondary coding for the content analysis.

References

[1] Khlood Ahmad, Mohamed Abdelrazek, Chetan Arora, Muneera Bano, and John Grundy. 2023. Requirements practices and gaps when engineering human-centered artificial intelligence systems. *Applied Soft Computing* 143 (2023), 110421.

[2] Laith Alzubaidi, Aiman Al-Sabaawi, Jinshuai Bai, Ammar Dukhan, Ahmed H. Alkenani, Ahmed Al-Asadi, Haider A. Alwzwazy, Mohamed Manoufali, Mohammed A. Fadhel, A. S. Albahri, et al. 2023. Towards risk-free trustworthy artificial intelligence: Significance and requirements. *International Journal of Intelligent Systems* 2023, 1 (2023), 4459198.

[3] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems* 10, 2 (2020), 1–37.

[4] Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology* 8, 1 (2005), 19–32.

[5] Jesse Josua Benjamin, Christoph Kinkeldey, Claudia Müller-Birn, Tim Korjakow, and Eva-Maria Herbst. 2022. Explanation strategies as an empirical-analytical lens for socio-technical contextualization of machine learning interpretability. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–25.

[6] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 78–91.

[7] Martin Böckle, Kwaku Yeboah-Antwi, and Iana Kouris. 2021. Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 3–20.

[8] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 807–819.

[9] Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. 2023. 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*, 188–203.

[10] Wasja Brunotte, Alexander Specht, Larissa Chazette, and Kurt Schneider. 2023. Privacy explanations—A means to end-user trust. *Journal of Systems and Software* 195 (2023), 111545.

[11] Enrico Bunde, Daniel Eisenhardt, Daniel Sonntag, Hans-Jürgen Profitlich, and Christian Meske. 2023. Giving DIAnA more TIME—Guidance for the design of XAI-based medical decision support systems. In *Proceedings of the International Conference on Design Science Research in Information Systems and Technology*. Springer, 107–122.

[12] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S. Poole, and Jofish Kaye. 2023. Healthcare AI treatment decision support: Design principles to enhance clinician adoption and trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.

[13] Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: A review and design principles for explanation user interfaces. In *Human-Computer Interaction—INTERACT 2021: 18th IFIP TC 13 International Conference*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, Kori Inkpen (Eds.), Lecture Notes in Computer Science, Vol. 12933, Springer, 619–640.

[14] Douglas Cirqueira, Markus Helfert, and Marija Bezbradica. 2021. Towards design principles for user-centric explainable AI in fraud detection. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 21–40.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[16] Rachel Eardley, Sue Mackinnon, Emma L. Tonkin, Ewan Soubutts, Amid Ayobi, Jess Linington, Gregory J. L. Tourte, Zoe Banks Gross, David J. Bailey, Russell Knights, et al. 2022. A case study investigating a user-centred and expert informed 'companion guide' for a complex sensor-based platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–23.

[17] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

[18] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. Fostering human agency: A process for the design of user-centric XAI systems. In *Proceedings of the 41st International Conference on Information Systems*, 1–18. Retrieved from https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12

[19] Ilina Georgieva, Claudio Lazo, Tjerk Timan, and Anne Fleur van Veenstra. 2022. From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI and Ethics* 2, 4 (2022), 697–711.

[20] Hongyan Gu, Yuan Liang, Yifan Xu, Christopher Kazu Williams, Shino Magaki, Negar Khanlou, Harry Vinters, Zesheng Chen, Shuo Ni, Chunxu Yang, et al. 2023. Improving workflow integration with XPath: Design and evaluation

of a human-AI diagnosis system in pathology. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–37.

[21] A. K. M. Bahalul Haque, A. K. M. Najmul Islam, and Patrick Mikalef. 2023. Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change* 186 (2023), 122120.

[22] Xin He, Yeyi Hong, Xi Zheng, and Yong Zhang. 2023. What are the users' needs? Design of a user-centered explainable artificial intelligence diagnostic system. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1519–1542.

[23] Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, and Christian Janiesch. 2023. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management* 69 (2023), 102538.

[24] Katja Herrmanny and Helma Torkamaan. 2021. Towards a user integration framework for personal health decision support and recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 65–76.

[25] Yan Jia, John McDermid, Nathan Hughes, Mark Sujan, Tom Lawton, and Ibrahim Habli. 2023. The need for the human-centred explanation for ML-based clinical decision support systems. In *Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. IEEE, 446–452.

[26] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. 2021. EUCA: A practical prototyping framework towards end-user-centered explainable artificial intelligence. arXiv:2102.02437. Retrieved from https://arxiv.org/abs/2102.02437

[27] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. 2024. Do stakeholder needs differ?—Designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. *International Journal of Human-Computer Studies* 181 (2024), 103160.

[28] Soyeon Kim, Junho Choi, Yeji Choi, Subeen Lee, Artyom Stitsyuk, Minkyoung Park, Seongyeop Jeong, You-Hyun Baek, and Jaesik Choi. 2023. Explainable AI-based interface system for weather forecasting model. In *Proceedings of the International Conference on Human-Computer Interaction*. Springer, 101–119.

[29] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

[30] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137.

[31] Samuli Laato, Miika Tiainen, A. K. M. Najmul Islam, and Matti Mäntymäki. 2022. How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research* 32, 7 (2022), 1–31.

[32] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. Retrieved from http://www.jstor.org/stable/2529310

[33] Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Meaningful explanation effect on user's trust in an AI medical system: Designing explanations for non-expert users. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–39.

[34] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez I Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[35] Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, et al. 2021. FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv:2109.09658. Retrieved from https://arxiv.org/abs/2109.09658

[36] Duri Long, Mikhail Jacob, and Brian Magerko. 2019. Designing co-creative AI for public spaces. In *Proceedings of the 2019 Conference on Creativity and Cognition*, 271–284.

[37] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

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

[39] Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. 2021. Principles of explanation in human-AI systems. arXiv:2102.04972. Retrieved from https://arxiv.org/abs/2102.04972

[40] Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology* 18 (2018), 1–7.

[41] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Explainable recommendation: When design meets trust calibration. *World Wide Web* 24, 5 (2021), 1857–1884.

[42] Luis Oberste, Florian Rüffer, Okan Aydingül, Johann Rink, and Armin Heinzl. 2023. Designing user-centric explanations for medical imaging with informed machine learning. In *Proceedings of the International Conference on Design Science Research in Information Systems and Technology*. Springer, 470–484.

[43] Nicola Palladino. 2021. Filling the gap between principle and practice: Building an ethical and human rights-based tool-kit for AI development. *GIGANET Annual Symposium*

[44] Michael Ridley. 2023. Using folk theories of recommender systems to inform human-centered explainable AI (HCXAI). *The Canadian Journal of Information and Library Science* 46, 2 (2023), 1–19.

[45] Ute Schmid and Britta Wrede. 2022. What is missing in XAI so far? an interdisciplinary perspective. *KI—Künstliche Intelligenz* 36, 3 (2022), 303–315.

[46] Tjeerd A. J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel Van Den Bosch. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies* 154 (2021), 102684.

[47] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67.

[48] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D. J. Peters, Tanya Horsley, Laura Weeks, et al. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine* 169, 7 (2018), 467–473.

[49] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Received 3 June 2024; revised 16 September 2025; accepted 1 October 2025