UNIVERSITY OF TURKU **ABSTRACT**

	Bachelor's thesis
x	Master's thesis
	Licentiate's thesis
	Doctoral dissertation

Subject	Information Systems Science	Date	18.3.2021
Author	Miika Tiainen	Number of pages	86+appendices
Title	To whom to explain and what? Systematic literature review on empirical studies on Explainable Artificial Intelligence (XAI)		
Supervisors	D.Sc. (Econ. & Bus. Adm.) Matti Mäntymäki, MSc. Samuli Laato		

Expectations towards artificial intelligence (AI) have risen continuously because of machine learning models' evolution. However, the models' decisions are often not intuitively understandable. For this reason, the field of Explainable AI (XAI) has emerged, which tries to create different techniques to help users understand AI better. As AI's use spreads more broadly in society, it becomes like a co-worker that people need to understand. For this reason, AI-human interaction in research is of broad and current interest.

This thesis outlines the current empirical XAI research literature themes from the human-computer interaction (HCI) perspective. This study's method is an explorative, systematic literature review carried out following the PRISMA (Preferred Research Items for Systematic Reviews) method. In total, 29 articles that concluded an empirical study into XAI from the HCI perspective were included in the review. The material was collected based on database searches and snowball sampling. The articles were analyzed based on their descriptive statistics, stakeholder groups, research questions, and theoretical approaches. This study aims to determine what factors made users consider XAI transparent, explainable, or trustworthy and to whom the XAI research was intended.

Based on the analysis, three stakeholder groups to whom the current XAI literature was aimed for emerged: end-users, domain experts, and developers. This study's findings show that domain experts' needs towards XAI vary greatly between domains, whereas developers need better tools to create XAI systems. The end-users, on their part, considered case-based explanations unfair and wanted to have explanations that "speak their language". Also, the results indicate that the effect of current XAI solutions on users' trust towards AI systems is relatively small or even non-existing. The studies' direct theoretical contributions and the number of theoretical lenses used were both found out to be relatively low.

This thesis's most immense contribution is to provide a synthesis of the extant empirical XAI literature from the HCI perspective, which previous studies have rarely brought together. Continuing this thesis, researchers can further investigate research avenues such as explanation quality methodologies, algorithm auditing methods, users' mental models, and prior conceptions about AI.

Key words	artificial intelligence, explainable AI, machine learning, human-computer interaction

TURUN YLIOPISTO

TIIVISTELMÄ

	Kandidaatintutkielma
x	Pro gradu -tutkielma
	Lisensiaatintutkielma
	Väitöskirja

Oppiaine	Tietojärjestelmätiede	Päivämäärä	18.3.2021
Tekijä	Miika Tiainen	Sivumäärä	86+liitteet
Otsikko	Mitä selitetään ja kenelle? Systemaattinen kirjallisuuskatsaus empiirisiin tutkimuksiin selittävästä tekoälystä (XAI)		
Ohjaajat	KTT Matti Mäntymäki, DI Samuli Laato		

Odotukset tekoälyä kohtaan ovat kohonneet jatkuvasti koneoppimismallien kehittymisen vuoksi. Mallien tekemät päätökset eivät usein ole ihmiskäyttäjälle vaistonvaraisesti ymmärrettävissä. Tätä ongelmaa ratkomaan on syntynyt selittävän tekoälyn tutkimuskenttä, joka luo erilaisia tekniikoita käyttäjien ymmärryksen tueksi. Kun tekoälyn käyttö yhteiskunnassa yleistyy laajemmin, tulee siitä ikään kuin työkaveri, jota ihmisten tulee ymmärtää. Tästä syystä tekoälyn ja ihmisen välisen vuorovaikutuksen tutkiminen on nyt laajan mielenkiinnon kohteena.

Tässä pro gradu -tutkielmassa hahmotellaan selittävän tekoälyn tutkimuskentän ajankohtaisia teemoja, ihmisen ja tietokoneen välisen vuorovaikutuksen näkökulmasta. Tutkielman metodi on tutkiva, systemaattinen kirjallisuuskatsaus, ja se suoritettiin seuraten PRISMA-ohjeistusta. Katsaukseen valikoitui yhteensä 29 ihmisen ja tietokoneen vuorovaikutuksen näkökulmasta selittävää tekoälyä empiirisesti tutkinutta artikkelia. Aineisto kerättiin tietokantahakujen ja lumipallo-otannan avulla. Tutkimuksia eriteltiin artikkeleja kuvailevien tietojen, niiden kohdeyleisön, tutkimuskysymysten sekä teoreettisten lähestymistapojen kautta. Tutkielman tarkoituksena on selvittää, millaiset tekijät saivat käyttäjät pitämään tekoälyä läpinäkyvänä, selitettävissä olevana tai luotettavana, sekä kenelle aihepiirin tutkimus oli suunnattu.

Analyysin perusteella löytyi kolme ryhmää, joille nykyistä kirjallisuutta on suunnattu: loppukäyttäjät, toimialojen asiantuntijat sekä tekoälyn kehittäjät. Tutkielman tulokset osoittavat, että asiantuntijoiden tarpeet selittävää tekoälyä kohtaan vaihtelevat laajasti toimialojen välillä, kun taas sen kehittäjät kaipaisivat parempia työkaluja tuekseen. Loppukäyttäjien havaittiin pitävän tekoälyn antamia tapauskohtaisia esimerkkejä epäreiluina, ja haluavan juuri heitä puhuttelevia selityksiä. Tulokset ilmaisevat, että nykyisten selittävien tekoälytekniikoiden vaikutukset käyttäjien luottamukseen tekoälyä kohtaan ovat vähäisiä. Tutkimusten tieteellisen panosten ja niiden käyttämien teoreettisten näkökulmien määrän havaittiin olevan suhteellisen pieniä.

Tämän tutkielman suurin tieteellinen panos on luoda yhteenveto empiiriseen, selittävän tekoälyn tutkimuskirjallisuuteen, ihmisen ja tietokoneen välisen vuorovaikutuksen näkökulmasta. Tätä näkökulmaa aiempi kirjallisuus on vain harvoin saattanut kokoon. Tutkielma avaa useita näkymiä jatkotutkimukselle, esimerkiksi selitysten laatumetodien, algoritmien auditointimenetelmien, käyttäjien ajatusmallien sekä aiempien käsitysten vaikutusten näkökulmista.

Avainsanat	tekoäly, selittävä tekoäly, koneoppiminen, ihmisen ja tietokoneen vuorovaikutus

UNIVERSITY OF TURKU
Turku School of Economics

TO WHOM TO EXPLAIN AND WHAT?

Systematic literature review on empirical studies on Explainable Artificial Intelligence (XAI)

Master's Thesis in Information Systems Science

Author
Miika Tiainen

Supervisors
D.Sc. (Econ. & Bus. Adm.) Matti Mäntymäki
MSc. Samuli Laato

18.3.2021
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Table of contents

List of figures

List of tables

1 INTRODUCTION

While artificial intelligence (AI) research goes back to decades ago, it has been the subject of extensive public debate during the 2010s and especially early 2020s – one could even say there is AI hype (Dwivedi et al. 2021). Expectations towards it have risen continuously because of the recent evolvement of machine learning (ML) techniques, deep learning, and neural networks. Unlike easily explainable, rule-based AI systems, machine learning models are black boxes that need to be explained via post hoc analysis (Arrieta et al., 2020). Furthermore, as AI can be used as a support system in high-stakes decision-making, its explainability is not only advisable but mandatory so that informed decisions can be made (Duan et al. 2019; Wang et al. 2019). As humans interact with AI more and more, it becomes like a co-worker or a friend that people need to understand. For this reason, AI-human interaction in research is of broad and current interest. In this research, an explorative, systematic literature review on empirical research on XAI from the human-computer interaction (HCI) perspective is conducted to map the discussions and trends in the field regarding user's experience on AI.

Simultaneously, AI systems have become so sophisticated that human guiding during its operation process is barely needed after the ML model has been developed. This automation also has its flip sides: many fears and challenges are associated with the increased use of AI (Dwivedi et al. 2021). For example, The World Economic Forum (Bossmann 2016) has identified nine major, generalized ethical issues and threats that the increasing use of artificial intelligence can pose: push people to unemployment, increase inequality, erode humanity, cause artificial stupidity, increase racism, cause security issues, and unforeseen consequences, result in the loss of the human control and the unclear legal status of AI systems.

As AI's use across society becomes more widespread, there is a need to understand its decisions and the logic behind them better (see, e.g., Došilović et al. 2018; Hagras 2018). Processes of how and why AI makes decisions and predictions should be more transparent and explainable so that even ordinary people can understand and trust them. In other words, the user should recieve more information and guidance than only the output of the model (Hagras 2018). Virtually every article discussing the transparency of AI names the black box phenomenon in ML as the biggest obstacle for the increasing use of artificial intelligence (see, e.g., Samek et al. 2017; Adadi and Berrada 2018; Došilović et al. 2018; Barredo Arrieta et al. 2020; Weitz et al. 2020). In this context, the term "black-box" refers to situations where human abilities cannot deduce why artificial intelligence has come to a particular conclusion because the systems are learning new from the data.

In this study, the focus is on the machine learning models rather than AI generally. However, the model is just one element to make AI explainable; developers must combine it with other data based on the audience's needs and expectations (Ferreira and Monteiro

2020). Besides increasing users' trust, there are other product goals explainability may help to fulfill. It facilitates verification and improvement of the system itself, learning from the system, and compliance to legislation (Samek et al. 2017). With increasing regulation, the pursuit of explainability is already somewhat mandatory. In the European Union, based on the General Data Protection Regulation, the subject of a decision made by artificial intelligence is entitled to an explanation of why the system ended up in such a decision (Goodman and Flaxman 2017).

The purpose of this thesis is to provide an overview of the existing XAI (explainable artificial intelligence) literature from the human-computer interaction (HCI) viewpoint through a systematic, conceptual literature review, following the PRISMA systematic literature review guidelines.

This research is exploratory; the goal is to find what kind of user-centric XAI literature there is and what are the main discussions in the field. The aim is to shed light on those topics and possibly identify research gaps that researchers have not yet discussed. In this research, explainable AI is defined as the pursuit to provide information and reasons about the AI's functioning to a particular audience (adapted from Barredo Arrieta et al. 2020). The HCI viewpoint means that this thesis focuses on examining the perspective of users. In this context, the user indicates a person who uses AI in work or is the subject of a decision made by AI, such as dealing with a government agency. An analysis of different user groups is provided later in this thesis.

Based on the literature review, the essential findings, and their implications, potential future research avenues are discussed. It is determined that the current literature presents many different technical solutions to increase explainability. However, the user's point of view (what makes artificial intelligence explainable in their opinion) has been less emphasized – this is the research gap on which this thesis focuses.

1.1 Defining research field

1.1.1 Defining explainable AI

XAI is a relatively new research field that has amassed broader scientific and societal interest during the second half of the 2010s. The application areas in which XAI is being used or discussed vary from commodities to potentially life-saving applications, and they are continually evolving. Examples included medical diagnostics tools (see, e.g., Bussone et al. 2015; Xie et al. 2019; Thomas and Haertling 2020), autonomous systems, such as cars (see, e.g., Robb et al. 2019; Li et al. 2020), banking operations (see, e.g., Cirqueira et al. 2020; van den Berg and Kuiper 2020), recommender systems (see, e.g., Samih et al.

2019; Ngo et al. 2020) education systems (see, e.g., Holstein et al. 2019; Putnam and Conati 2019) decisions of public authorities (see, e.g., Cheng et al. 2019; Deeks 2019), and even agricultural applications (see Kenny et al. 2019), to name some.

The number of studies on XAI has increased dramatically (see, e.g., Barredo Arrieta et al. 2019; Ferreira and Monteiro 2020). Before 2017, the research in the field concentrated mainly on the concept of interpretable AI. The increase in the number of studies reflects the growing need for AI knowledge in the scientific community and society. In addition to interpreting AI and predicting its decisions (interpretability), there is a need to explain how and through what process its decisions are made (explainability). There are various other closely related words used to describe explainability besides interpretability, such as accountability, transparency, responsibility, understandability, etc. (see, e.g., Brennen 2020; van den Berg and Kuiper 2020). The diversity of terms has hampered the emergence of unified terminology (Brennen 2020). The bumpy use of words might be due to the relative freshness of the research field and the fact that new research in the area is emerging rapidly worldwide.

Due to its diversity, there have been many different yet complementary pursues to define XAI as a phenomenon. For example, Gunning (2017) has identified XAI as the pursuit to "create a suite of machine learning techniques" that will help people to "understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." This definition sees XAI as a quest to provide understanding and trust to users and management practice. Additionally, the definition sees practitioners as active agents trying to provide explainability to passive users to help them understand.

By contrast, Barredo Arrieta et al. (2020) provide the following definition: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand." This definition thus shares the perspective of Gunning: XAI is needed to provide understanding. The difference is that Barredo Arrieta et al. (2020) approaches XAI as having multiple, different audiences that have various kinds of needs for explanations instead of management practice. Another difference is that in this definition, XAI is seen as an active entity itself, being capable of producing details and reasons. Thus, this definition manages to capture the current, self-developing nature of XAI. Therefore, in this thesis, XAI is approached from the perspective Barredo Arrieta et al. (2020) defined – XAI always stems from an audience's needs, and the viewpoint of users is at the center of this study.

1.1.2 Why is XAI needed?

Among the literature, multiple reasons are defined why explainability of ML models is needed. Adadi and Berrada (2018) propose at least four reasons or motivations from

which the need for XAI may stem. The four motivations are explaining to justify (justifications for a particular outcome), explaining to control (preventing thins from taking the wrong track, explaining to improve (explainability will make improving the models easier in the future), and explain to discover (to gain new knowledge). Explainability pursues are thus focused not only on enhancing users' understanding but also on fulfilling many other product goals.

XAI is especially needed in the fields that involve high-stakes decision-making, such as criminal justice and clinical decisions (see, e.g., Cai et al. 2019; Ferreira and Monteiro 2021). It could be said that explainability is paramount and a requirement so that AI can be used in these contexts. The high-stakes decision-makers need detailed and understandable explanations to humans and access to background data and depictions about the model's inner workings (see, e.g., Cai et al. 2019; Wang et al. 2019). It is also worth investing in everyday applications' explainability, such as remote home control systems. However, the explanations they require are at a more superficial level. The discussion of what are valuable explanations is at the heart of the HCI perspective. This discussion is also closely related to concepts from social sciences, such as philosophy and psychology.

1.1.3 Factors shaping XAI and its use

While the XAI as a research field has boomed during the last ten years, e.g., Holzinger (2018), Preece (2018), and Longo et al. (2020) point out that user experience and explainability within it is one of the oldest fields of study in computer science. They also highlight that early AI was interpretable and retraceable. Therefore, it was not reasonable to limit the literature review to only those articles published in recent years. However, it is worth noting that the rule-based imperative AI systems are very easily explainable. But with ML models, the system is continuously learning and its predictions evolving, so understanding these systems will require post-hoc analysis. Thus, explainability is no longer anymore a given. (Barredo Arrieta et al. 2019)

Figure 1 (below) shows the dilemma between performance and transparency in ML techniques (adapted from Došilović et al. 2018). Whereas the rule- and tree-based models are easily explainable, they lack the performance-abilities that the black-box models, such as artificial neural networks (ANN), have. On the same theme, Barredo Arrieta et al. (2019) identify two different sub-classifications for explainability in recent literature. Firstly, machine learning models can be somewhat transparent and interpretable. Secondly, developers can use XAI techniques after the machine learning model is trained (post-hoc) to make it more interpretable. Thus, explainability efforts can be made already at the models' design stage by choosing the more transparent methods or developing the AI afterward to improve human-computer interaction. The decisive question then is: are

the performance of the more transparent models sufficient for the chosen purpose? Often the answer is no. More effective methods are needed to seize the benefits of artificial intelligence, and post-hoc methods are then required to keep the human in the loop.

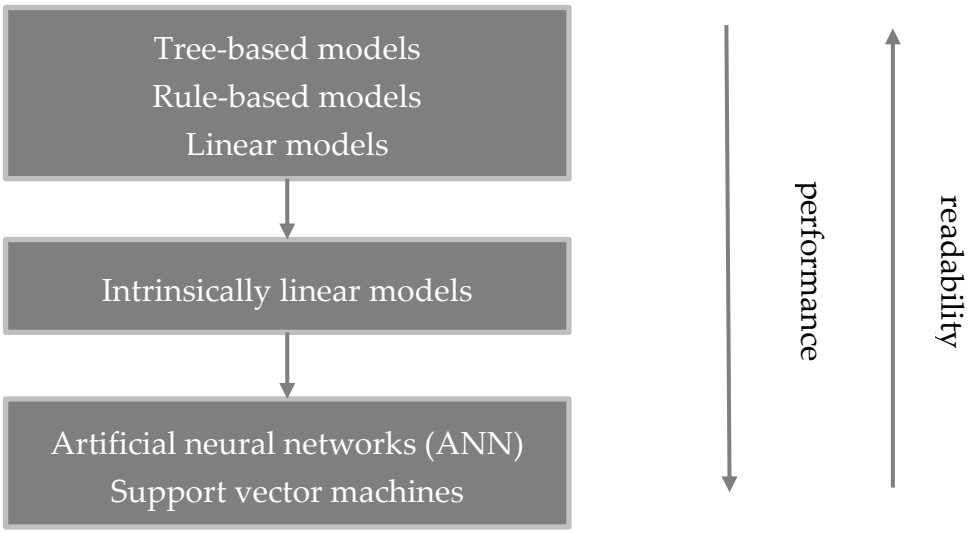

Figure 1 Performance-transparency trade-off (redrawn according to Došilović et al. 2018, 211)

It is worth noting that it is not always a matter of choosing the correct XAI technique. The implementation of XAI is also affected by other factors, ranging from the organization's capabilities to user receptivity and societal conditions. As an example, Chazette and Schneider (2020) present dimensions that "affect the elicitation and analysis of explainability" (pictured below in Figure 2). They state that users' needs and expectations towards AI may vary. Cultural and corporate values may affect its design and deployment. Laws and norms impose constraints on its development process. The design project itself has constraints since different organizations have alternating resources for their use.

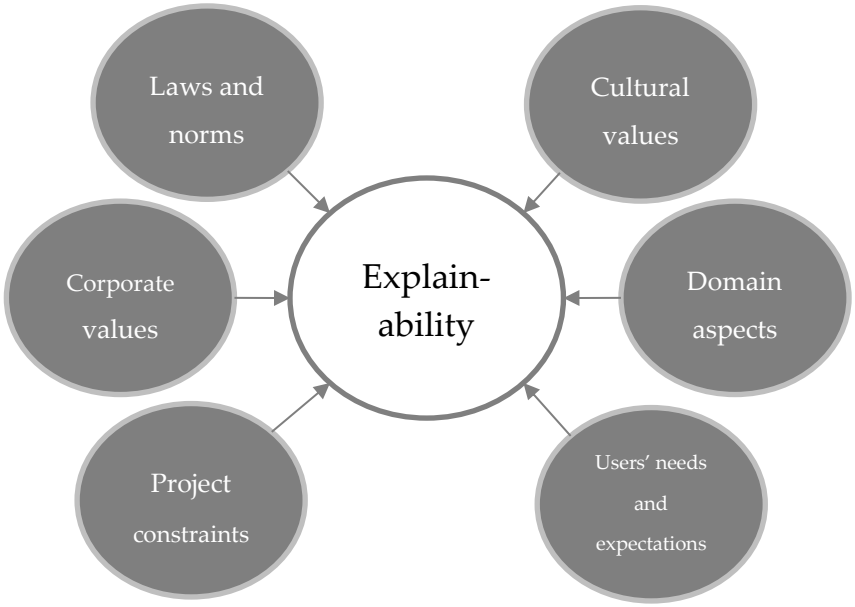

Figure 2 Dimensions shaping explainability (redrawn according to Chazette and Schneider 2020, 14)

Regarding these dimensions, research fields are just beginning to take shape. For example, with the coming of new regulation, such as GDPR, there is already some research and discussion on the impact of laws and standards on XAI (see, e.g., Doshi-Velez et al. 2017; Deeks 2019; Hacker et al. 2020). Some studies have also discussed the implications of project constraints, corporate resources, and values to the development of XAI (see, e.g., Holstein et al. 2019; Hong et al. 2020; Liao et al. 2020). The user needs and perspectives have been studied, but the field is significantly fragmented due to the diversity of needs.

Based on the familiarization with the existing literature, the research efforts have mainly focused on presenting technical explanatory solutions, emphasizing post-hoc methods. That is why the HCI perspective would be needed to see whether users consider those solutions as transparent or explainable and to what extent. The lack of user-centric studies and an emphasis on technical solutions have also been noted in many previous studies (see, e.g., Kirsch 2018; Narayanan et al. 2018; Gunning et al. 2019; Liao et al. 2020), so it is a reasonable starting point for this thesis as well.

1.2 Research questions

As a summary of the research field, there is a lack of comprehensive, user-centric literature on XAI and its potential understandability, trustworthiness, and explainability. Based on the research area and the complexity, missing clarity, and rapid development of XAI research in the HCI field, this thesis's research question is as follows:

How do users' perceptions of transparency, explainability, and trustworthiness of AI manifest in the HCI literature?

In support of the leading research question, additional research questions are used, which are the following:

- To whom is the XAI for? What are its stakeholders?
- What factors make AI transparent, explainable, or trustworthy for these target audiences?
- Do different audiences differ in their perceptions of the explainability of AI?
- What kind of different needs target audience groups have for explainability?

1.3 Structure of the study

The rest of the paper is structured as follows. First, the methods used in this study are described. The criteria used to include articles into the review and exclude irrelevant ones is stated, as well as how the collection process advanced. After that, a step-by-step overview of how the synthesis was conducted is provided.

Second, the findings of the study are presented. A descriptive analysis of the studies reviewed will be included, followed by a synthesis of the key findings based on stakeholder groups. After that, the theoretical lenses used in the studies to conduct their empirical research are discussed. The research questions of the articles are then grouped, and the findings are analyzed based on question-type.

Third, the research's key findings are summarized, and the broader implications of the findings are discussed. The limitations of this study's process and biases and limitations across and within the sample's studies are discussed. Limitations are followed by a section discussing future research avenues identified based on this study and the sample studies. In the last segment of this study, the concluding remarks are made.

2 RESEARCH MATERIALS AND METHODS

2.1 Study design

The research method of this thesis is a systematic, conceptual literature review. The review is carried out by following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method and essentially following the Prisma 2009 Checklist (Moher et al. 2009). It is worth noting that several other guidelines for conducting systematic literature reviews exist, such as Procedures for Performing Systematic Reviews (Kitchenham 2004) or the expectations for a review article by Webster and Watson (2002). However, Prisma is a rigorous and well-tested approach, the preferred SLR method in many journals (PRISMA 2021). It was chosen and seen suitable for this study because it is widely used in information systems science and even business (see, e.g., Mardani et al. 2018; Satalkina and Steiner 2020; Wieringa 2020).

The systematic literature review aims to create a complete and comprehensive description of the problem's treatment in the previous literature. Every step of the review will be described as accurately as possible to preserve its replicability. The author of this thesis is acquainted with the principles of research ethics and has carried out this research following good scientific practice (see, e.g., Finnish Advisory Board on Research Integrity 2012; University of Turku 2013).

The literature review process follows the flow chart of information created for the PRISMA method by Moher et al. (2009). The process is pictured in Figure 3 below. The first step was to combine the results from the selected search databases obtained with suitable keywords and their combinations. After the records were combined and duplicates removed, the documents were screened for relevance. Irrelevant articles for this study's scope were then removed based on screening through the studies' abstracts. The remaining pieces were then read through, some in their entirety, some only screened through to eliminate unnecessary articles for the review. The criteria for article eligibility to the next stage is stated in section 2.2. A qualitative synthesis was then done to draw the findings together from the studies. The synthesis was done by integrating general article data, findings of the articles, and future research areas to answer the research questions identified in section 1.2.

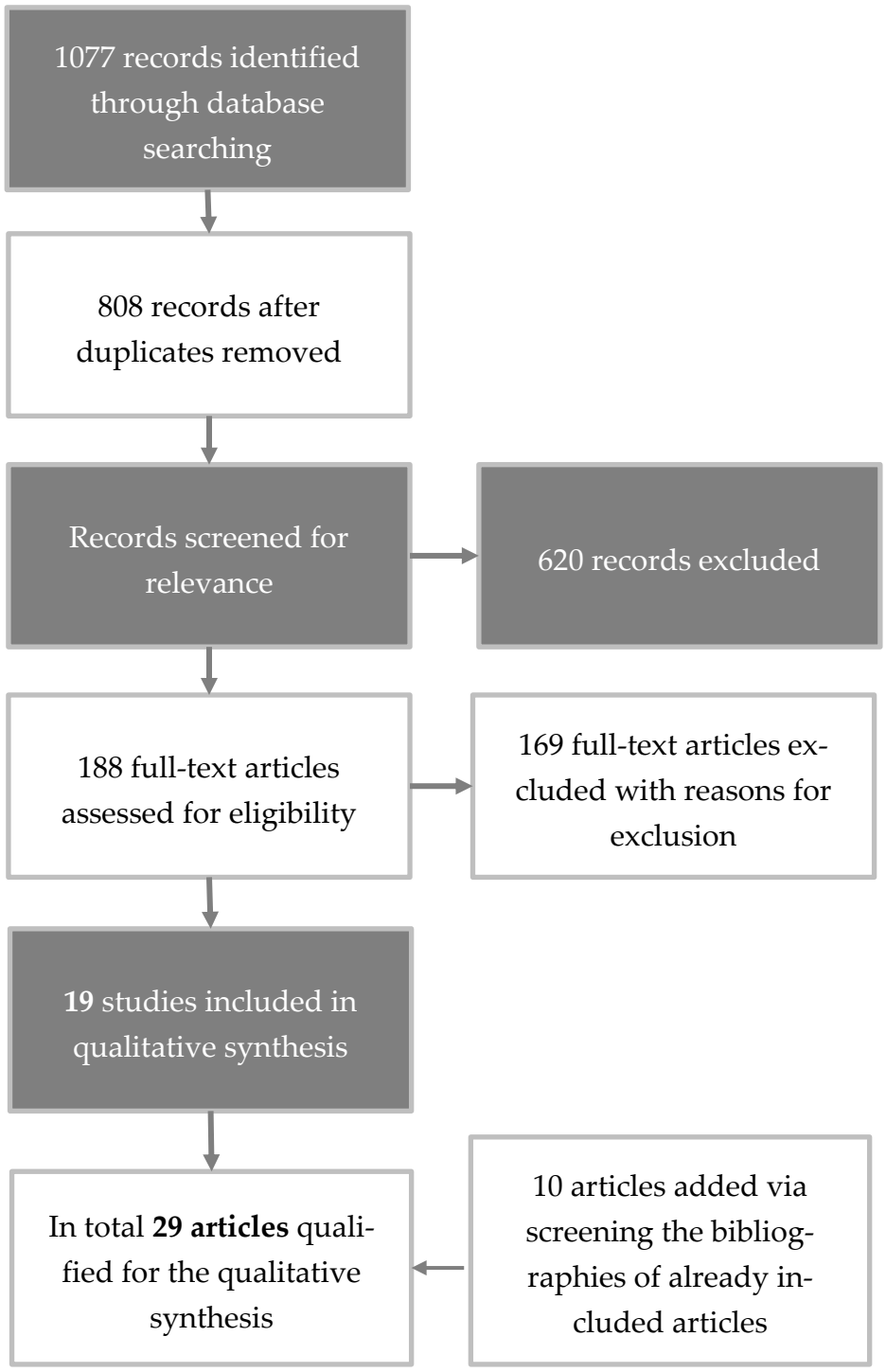

Figure 3 Flow of information through the phases of the systematic review (adapted from Moher et al. 2009)

2.2 Data collection

In this section, the grounds on which ineligible articles were excluded during the search process are first described. This is followed by the criteria by which the articles in the final phase were selected for inclusion in the literature review. In the second sub-section, the actual collection process through the databases – in other words – how the criteria were put into practice. The third sub-section provides information on how additional articles outside the reach of the database searches were identified and added to the review sample.

2.2.1 Inclusion and exclusion criteria of the articles

The articles to review were obtained from major scientific databases, and additional collection efforts (snowball sampling), the process itself is depicted in the following section. The first exclusion criterion was applied when creating the search chains: studies not in English were left out. The second criterion was applied immediately after the searches were conducted: those studies that were not related to AI were left out. Examples of unrelated articles are presented in the following section.

A two-phase inclusion and exclusion criterion was then applied to the remaining articles. First, a report was classified for further review, i.e., of interest, if at least one of the three conditions were met:

1. the abstract clearly stated that the article studies users' experience or perceptions about artificial intelligence from the perspective of explainable AI or
2. the abstract presented a particular explainability (post-hoc) solution and discussed its effectivity or
3. the abstract was very general, concise, or vague, and it could not be ruled out that it would not have any human-computer interaction related or user-centric sections.

In the second phase, the article would qualify for the actual synthesis if it would meet the following criterion:

1. the article presented results of an empirical study, providing qualitative or quantitative data describing users' attitudes, expectations, or requirements towards AI and its explainability, or its synonym.

An article was excluded if its empirical results only described a particular AI solution's usefulness relative to the starting point or some other solution. This exclusion was done to limit the number of articles. The inclusion and exclusion process is depicted in Figure

4 below. How the criteria expressed in this section was applied into practice is described next in the following sub-section.

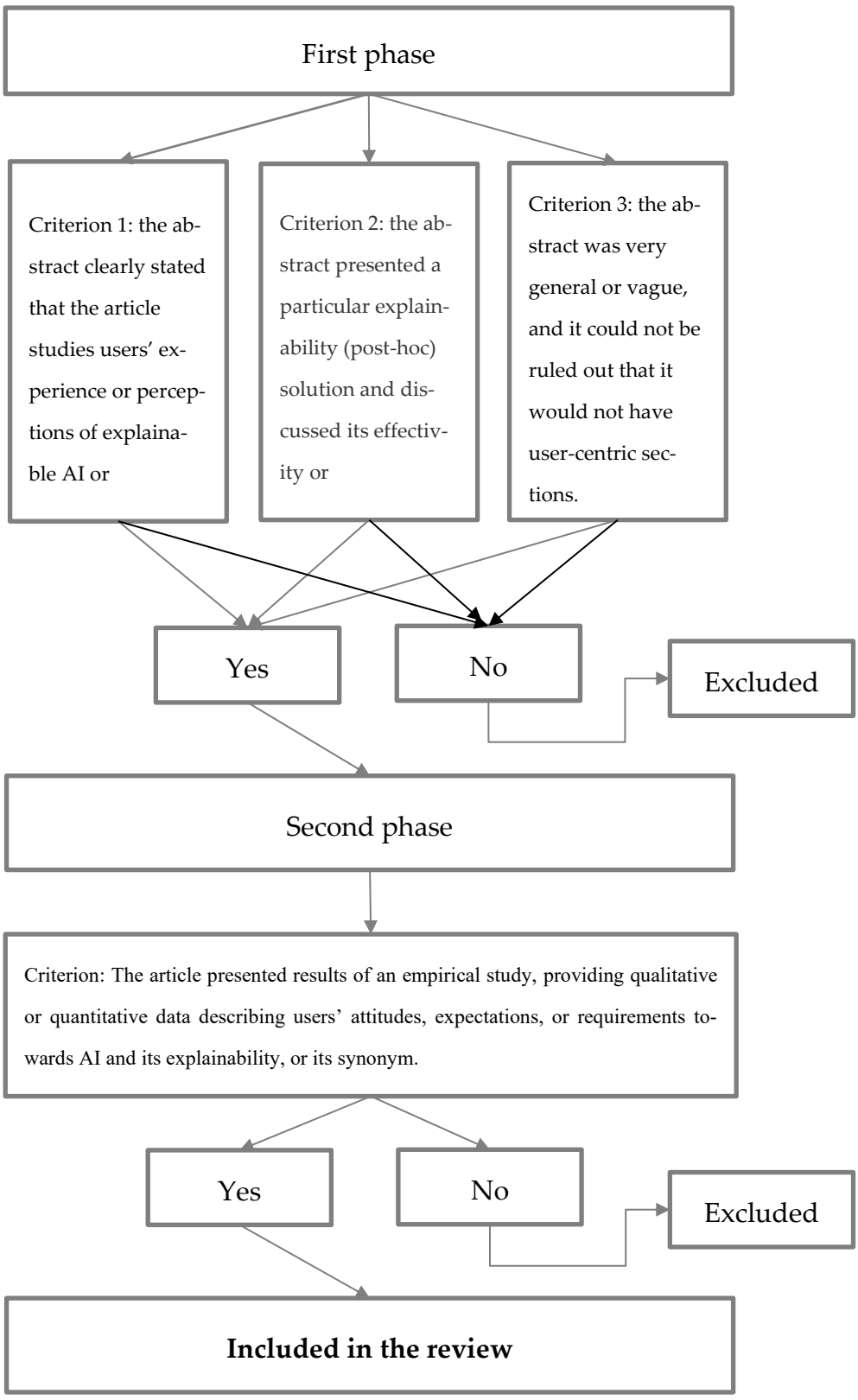

Figure 4 Inclusion and exclusion process of the articles

2.2.2 Articles collected through the research database search

This study reviews the search results from two academic databases: Scopus and Web of Science. Scopus was chosen because it indexes many other possibly essential databases, such as ACM, Springer, IEEE, and the DBLP Computer Science Bibliography. For reliability, the Web of Science was chosen to complement Scopus to ensure the broad coverage of search results. The results were later supplemented with articles collected through the bibliographies of the initially identified works. The searches to the databases targeted article titles, abstracts, and keywords. Due to the differences between the two search engines, the search strings used are listed in Figures 5 and 6. The keywords to the search were decided based on previous research and discussions with senior researchers. A preliminary searching of relevant articles was done. It was discovered that the terms XAI, transparent AI, interpretable AI, and accountable AI were all used to describe the research field.

TITLE-ABS-KEY(xai OR "Explainable AI" OR "transparent AI" OR "interpretable AI" OR "accountable AI" OR "AI explainability" OR "AI transparency" OR "AI accountability" OR "AI interpretability") AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"ed") OR LIMIT-TO (DOCTYPE,"bk") OR LIMIT-TO (DOCTYPE,"er") OR LIMIT-TO (DOCTYPE,"le") OR LIMIT-TO (DOCTYPE,"no")) AND (LIMIT-TO (LANGUAGE,"English"))

Figure 5 Search string for Scopus

(TI=(xai OR "Explainable AI" OR "transparent AI" OR "interpretable AI" OR "accountable AI" OR "AI explainability" OR "AI transparency" OR "AI accountability" OR "AI interpretability") OR AK=(xai OR "Explainable AI" OR "transparent AI" OR "interpretable AI" OR "accountable AI" OR "AI explainability" OR "AI transparency" OR "AI accountability" OR "AI interpretability") OR AB=(xai OR "Explainable AI" OR "transparent AI" OR "interpretable AI" OR "accountable AI" OR "AI explainability" OR "AI transparency" OR "AI accountability" OR "AI interpretability")) AND ***DOCUMENT TYPES:*** *(Article OR Abstract of Published Item OR Proceedings Paper)*

Figure 6 Search string for Web of Science

The search of the articles in Scopus and Web of Science concluded on October 20, 2020. Any relevant articles published after that are thus not included in the review. The search for additional items through snowball sampling (depicted in section 2.2.3.) was conducted after that, but no newer articles were included in the review via snowballing.

After the database searches were conducted, all citations were then downloaded as .csv files, which were then imported to Google Sheets for automatic formatting and then downloaded to Excel and combined. The citation source, a link to the article, citation count, abstract, and author keywords were included in the file. Besides, the status of a paper after each screening step was recorded.

Through Scopus, the search resulted in 724 articles. One article appeared twice among the data, causing the number to drop to 723. From Web of Science, 353 articles were found. Of those, 268 articles were duplicates (already found from Scopus). The relevant 85 articles left were then combined with Scopus's ones, resulting in 808 articles. It is worth noting that of the 808 articles, only 110 were published before 2017, and only seven out of those 110 articles were related to artificial intelligence.

The title and abstract of all the 808 articles were read thoroughly. Of those, in total, 131 items could be scratched immediately in the first screening of the title and the abstract since they were utterly unrelated to artificial intelligence or its applications. In the excluded articles, there were studies from the fields of mathematics and chemistry in which "XAI" was mentioned as part of a formula. Also, several search results related to biological or geographical studies in or near the Mozambican city of Xai-Xai or articles involving the indigenous Xai'xais people in Canada.

All the remaining 677 articles dealt with AI in some way. The papers were also classified based on their discipline. In practice, for example, items assessing medical algorithms are classified under medicine, while those dealing with self-driving cars are classified under engineering.

The articles were then divided into two groups: papers of interest to the current research (focusing on human-computer interaction) and irrelevant ones based on the first phase exclusion criteria described in the previous section. Among the irrelevant articles, studies were providing general overviews about the XAI field or its recent trends (see, e.g., Došilović et al. 2018; Gilpin et al. 2018; Holzinger 2018; Schoenborn and Althoff 2019), and articles discussing purely technical solutions (see, e.g., Guidotti et al. 2019; Jha et al. 2019; Lundberg et al. 2020; Schaaf et al. 2019). Also, some excluded papers discussed miscellaneous, merely technical cases in which XAI methods were applied (see, e.g., Keneni et al. 2019; Marino et al. 2018; Olszewska 2019; Thomas and Haertling 2020). Of the 677 articles, 188 articles met the criteria stated above and were therefore selected for further review. The remaining 489 items were excluded from the study at this stage.

The 188 articles left were then read in their entirety. However, if reading the paper became clear that it had nothing to do with XAI from the HCI perspective, the article was put aside. The purpose was to get a complete grasp of the article's topic and its study design, so it could be defined, which would make the final cut to the qualitative synthesis. The items excluded in this point were found out to be either only descriptional (see, e.g.,

Gunning and Aha 2019; Hagras 2018) or provided, for example, depictions of use cases or mentioned the perspective of the users but did not conduct an empirical study (see, e.g., Stumpf 2019; Zhu et al. 2018). The articles that were depicting user studies that provided information only about the effectiveness of one particular XAI technique, for example, were also left out (see, e.g., Kuwajima et al. 2019; Ming et al. 2019).

Out of 188 articles, only 19 met the criterion and thus qualified for the synthesis via databases. It was expected that in this stage, the total number of items would still drop significantly, so this result did not come as a surprise.

2.2.3 Additional articles

After the relevant articles were collected, backward snowball sampling was used to collect additional interesting papers through the already existing sample. Backward snowball sampling means the practice of screening the bibliographies of the articles included in the review (Wohlin 2014). Snowball sampling was done to see if there would be any articles that did not come up in the search results and should be included in the synthesis.

If an article seemed interesting based on its title, it was marked down and later searched via Google Scholar or other databases. All the titles that seemed exciting and related to AI were then searched and looked through the way done with the original articles (depicted in section 2.2.1). All combined, the bibliographies encompassed in total 848 unique items, of which 17 were considered highly relevant for this research and fulfilled the first phase of the criteria. The same inclusion/exclusion criteria were applied for these articles to those previously identified via Scopus and Web of Science (depicted in section 2.2.1).

Combined with the articles found in the previous stage, the pool of papers for the synthesis was at this point in total 36. Then the 17 new articles were read more carefully, which showed that seven of the items could, nevertheless, be left out since their contributions did not match the perspective of this study when mirrored against the second phase of the exclusion criteria. Thus, the final total number of articles qualified for the sample would be 29 (the ten new pieces combined with previously identified 19 articles). One reason for the relatively high amount (17) of interesting new papers found via snowball sampling was that the search string did not include the term “machine learning”. Another reason is that some of the articles referred to words that are not in widespread use, such as “intelligibility” (see, e.g., Lim and Dey 2009; Lim et al. 2009). The process through this stage was pictured above in Figure 3.

How the articles were then analyzed is discussed in the next section 2.3., and the actual results in section 3.

2.3 Data analysis

This section discusses the data analysis process was conducted after the articles were collected. Reasons and motivations for the approaches chosen are also provided.

The synthesis process started by familiarizing with the articles by reading them repeatedly and taking notes about compelling factors. The author of this thesis read through all the papers; no additional researchers participated in the reading process. The process following the familiarization phase can be divided into four parts, described in the following paragraphs.

First, descriptive data from the studies were obtained. Publication year, venue, methods, sample, application area, and country of data origin were collected. This collection was done to provide a good overview of the literature and see if there were any patterns regarding, for example, methods used, geographical distribution, or if some journal or publication channel was heavily emphasized among the sample.

Second, the articles were sorted into clusters based on to whom XAI was aimed. This clustering was done to see what different stakeholder groups researchers considered relevant to XAI research and how the groups' roles were justified – basically to answer the first additional research question of this thesis. Understanding the perspectives of other researchers on the groups could help to unite the fragmented research field.

The studies were categorized into three clusters: developers, domain experts, and end-users. If the stakeholder group of the article was unclear, it was left aside and classified later. This unclarity was the case of three items in total, all of which were later grouped under end-user studies. The categorization process was iterative. Articles were combined into bigger groups until saturation was reached, and categories could not be combined anymore. For example, a fourth group, other stakeholders, was first identified. This category was a heterogeneous group spanning from the individuals who work in AI-related fields to corporate shareholders and community members: individuals affected by the AI but are not necessarily using or developing it themselves. However, this group was combined with the end-user group since only one study was in the category.

Third, the following steps were then taken to answer the other additional research questions and, eventually, to this thesis's central research question. As a third step, the theoretical underpinnings of the literature were looked at closely. The aim was to discover the most popular and promising theoretical approaches in understanding XAI among the sample articles. The goal was to find the theoretical lenses through which the empirical parts of the studies were conducted. Papers in which researchers had formulated research hypotheses were looked through, and the theoretical basis on which those hypotheses were formed was read closely. The collection of theoretical lenses could potentially help further endeavors, as future researchers could also use these theories to form their

theoretical basis and hypotheses. The research questions the studies used were collected and paired with similar questions among the studies, following the idea of "elementary question forms" formulated by Bunge (1967). This research question analysis was done to find the similarities among the studies' baselines and how similar their findings would be.

One type of analysis excluded at this stage because of the schedule was looking at the articles based on their application area. This analysis could have been done by grouping articles based on their research questions into groups investigating similar issues instead of or in addition to a question-based approach. This approach can still be utilized in further research based on this thesis.

Fourth and last, all the articles' key findings were collected and grouped based on the stakeholder clusters identified before. The aim was to find out what kind of themes the items discussed, how the discussed themes would differ between stakeholder groups and similarities in the particular audience group's findings. This step was essential to provide answers to the primary research question and the second additional research question. The whole data analysis process step by step is depicted below in Figure 7.

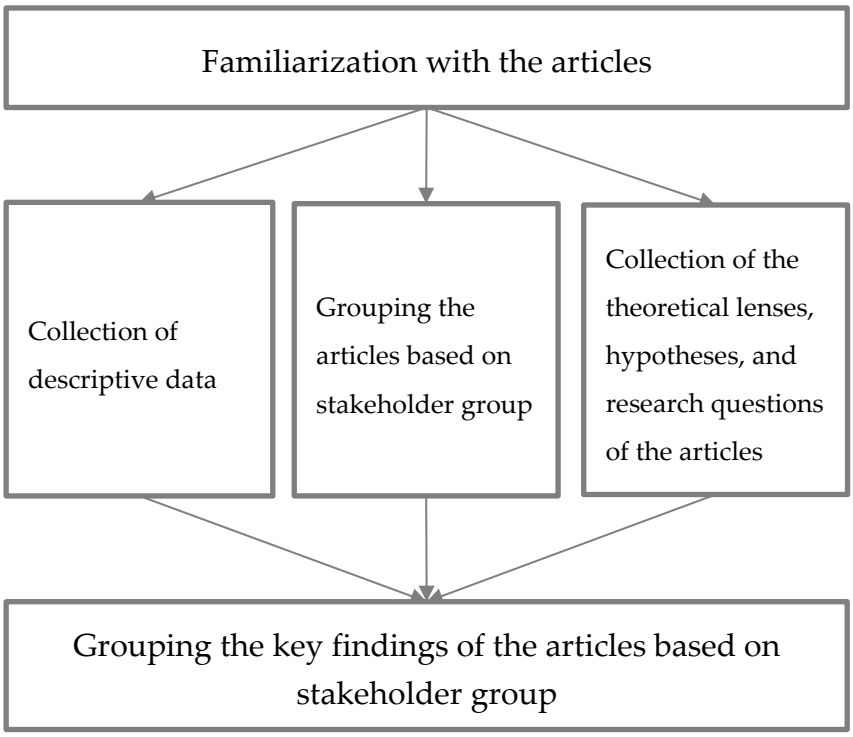

Figure 7 Flowchart of the data analysis process

3 FINDINGS

3.1 Descriptive information regarding the studies

In this section, a detailed look into the descriptive statistics of the sample articles is provided. Overview of the sample articles with descriptive details, such as application area, method, and sample, are provided below (see Table 1).

Article	Stakeholder group	Application area	Method	Sample	Country of data origin
Binns et al. 2018	end-users	algorithmic justice	qualitative and quantitative, user experiment, survey, and semi-structured interviews	19 end-users (lab study), 325 end-users (between-subjects study), 65 end-users (within-subjects study)	United Kingdom
Brennen 2020	end-users	AI explainability definition	qualitative, semi-structured interviews	40 stakeholders in interviews and 24 stakeholder participants in focus groups	not stated
Broekens et al. 2010	end-users	autonomous systems (cooking)	quantitative, user experiment, and survey	30 end-users in between-subject study	not stated
Bussone et al. 2015	domain experts	medical diagnostics	qualitative and quantitative, between-group study and interviews	Seven primary care practitioners and one nurse	not stated
Cai et al. 2019	domain experts	medical diagnostics	qualitative, structured interviews and user experiment	21 pathologists in lab study and interviews	not stated
Chazette & Schneider 2020	end-users	navigation systems	quantitative and qualitative, survey	107 end-users	Germany and Brazil

Cheng et al. 2019	end-users	university admissions	quantitative and qualitative, design workshop and survey	12 participants (university community members or future applicants) in design workshops and 199 end-users via MTurk in the survey	not stated
Cirqueira et al. 2020	domain experts	fraud detection (banking)	qualitative, semi-structured interviews	Three banking fraud specialists from one bank	Austria
Cramer et al. 2008	end-users	art recommender systems	qualitative and quantitative, user experiment, survey, and interview	60 end-users	not stated
Dodge et al. 2019	end-users	criminal justice	quantitative and qualitative, survey	160 end-users via MTurk	United States
Ehsan et al. 2019	end-users	gaming	quantitative and qualitative, user experiment and surveys	128 end-users via TurkPrime	United States, India
Eiband et al. 2018	end-users	fitness application	qualitative and quantitative, workshops, semi-structured interviews, card sorting, user experiment	14 end-users in interviews, 11 end-users in card sorting, 16 end-users in a user experiment	not stated
Eslami et al. 2018	end-users	online advertising	qualitative, user experiment and interviews	32 end-users	United States
Hohman et al. 2019	developers	visual explanations	qualitative and quantitative, interviews, survey, user experiment	12 data scientists from on technology company	not stated
Holstein et al. 2019	developers	commercial ML product teams	qualitative, semi-structured	35 practitioners in interviews from 10 companies, 267 ML	"multiple countries"

			interviews, survey	practitioner participants in the survey	
Hong et al. 2020	developers	ML systems' design processes	qualitative, semi-structured interviews	22 ML practitioners from 20 companies	not stated
Liao et al. 2020	developers	explainability design practices	qualitative, semi-structured interviews	20 UX and design practitioners from IBM	United States + two participants from unspecified countries
Lim and Dey 2009	end-users	context-aware applications	qualitative and quantitative, survey	250 end-users in the first survey, 610 end-users in the second (both via MTurk)	not stated
Lim et al. 2009	end-users	context-aware applications	quantitative and qualitative, user experiment and survey	53 end-users in the first experiment, 158 end-users in the second (both via MTurk)	not stated
Ngo et al. 2020	end-users	recommender systems	qualitative, semi-structured interviews	Ten end-users	not stated
Oh et al. 2018	end-users	human-AI art co-creation	qualitative and quantitative, user experiment, survey, and semi-structured interviews	30 end-users	not stated
Putnam & Conati 2019	end-users	tutoring systems	quantitative and qualitative, usability testing and survey	Nine university students	not stated
Schrills & Franke 2020	end-users	visual explanations	quantitative, user experiment	83 end-users	not stated
van der Waa et al. 2020	end-users	decision support systems	qualitative and quantitative,	40 end-users	not stated

			structured interviews, survey		
Wang et al. 2019	domain experts	medical diagnostics	qualitative, user experiment, interviews	14 medical professionals from one hospital	not stated
Weitz et al. 2019	end-users	speech recognition, virtual agents	quantitative, user experiment, and survey	30 end-users	not stated
Weitz et al. 2020	end-users	speech recognition, virtual agents	quantitative, user experiment, and survey	60 end-users	not stated
Xie et al. 2019	domain experts	medical diagnostics	qualitative, semi-structured interviews	Six medical professionals	United States (California)
Yin et al. 2019	end-users	speed dating	quantitative, user experiment	1994 end-users in the first experiment, 757 end-users in the second, 1042 in the third (all via MTurk)	United States

Table 1 Overview of the sample articles

Of the sample articles, a clear majority was published during 2018–2020, with 24 in total. The publishing years formed two clusters, another being during the years 2008-2010 with four articles. Apart from the two groups, only one study, Bussone et al. (2015), was published between them. This finding reflects the explosive growth of the research field in recent years. The temporal distribution of the articles is illustrated in Figure 8 below.

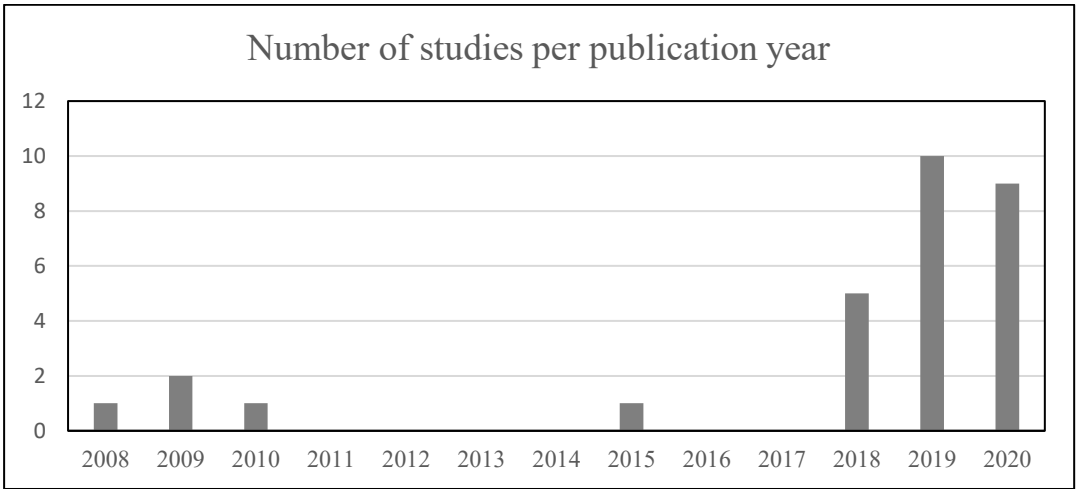

Figure 8 Number of the studies per publication year

Ten studies used qualitative methods, five studies used quantitative methods, and 13 articles collected qualitative and quantitative data regarding research methods. The most popular research methods among the sample were surveys and interviews; both were usually combined with user experiments. The overwhelming majority of the conducted interviews were semi-structured. Among the less used methods in the sample were workshops, card sorting, and between-group study. The studies' sample sizes varied from interviewing three domain experts to conducting three different user experiments in total 3 793 end-users via Amazon's crowdsourcing marketplace Mechanical Turk. The majority of the studies had from a few dozen to about one hundred participants. Of the interview studies, 13 out of 19 had a sample under 30 interviewees. Every article included in the domain expert or developer stakeholder group except Holstein et al. (2019) had under 30 interviewees. This number is understandable, as the interviews of practitioners and professionals in this sample were generally more homogenous and in-depth than ordinary end-users, and the saturation in them might thus be achieved earlier (see, e.g., Baker and Edwards 2012; Hennink et al. 2017). A few end-user studies had samples between nine and fifteen interviewees (Eiband et al. 2018, 14 fitness application users; Putnam and Conati 2019, nine university students; Ngo et al. 2020, ten Netflix users), which might affect their reliability. In the survey studies, in total, six studies had a sample under 150 participants. These studies all focused on some particular XAI system, which might explain the little bit lower sample size. The studies' sample sizes containing either user surveys or interviews are depicted in figures 9 and 10 below.

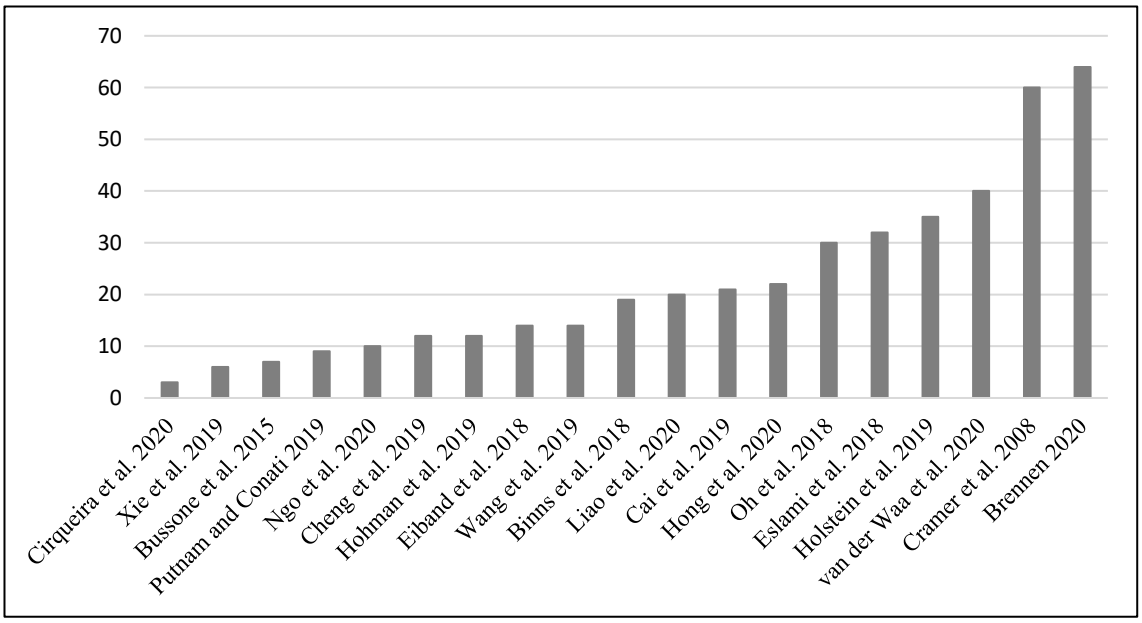

Figure 9 Sample sizes of the interview studies

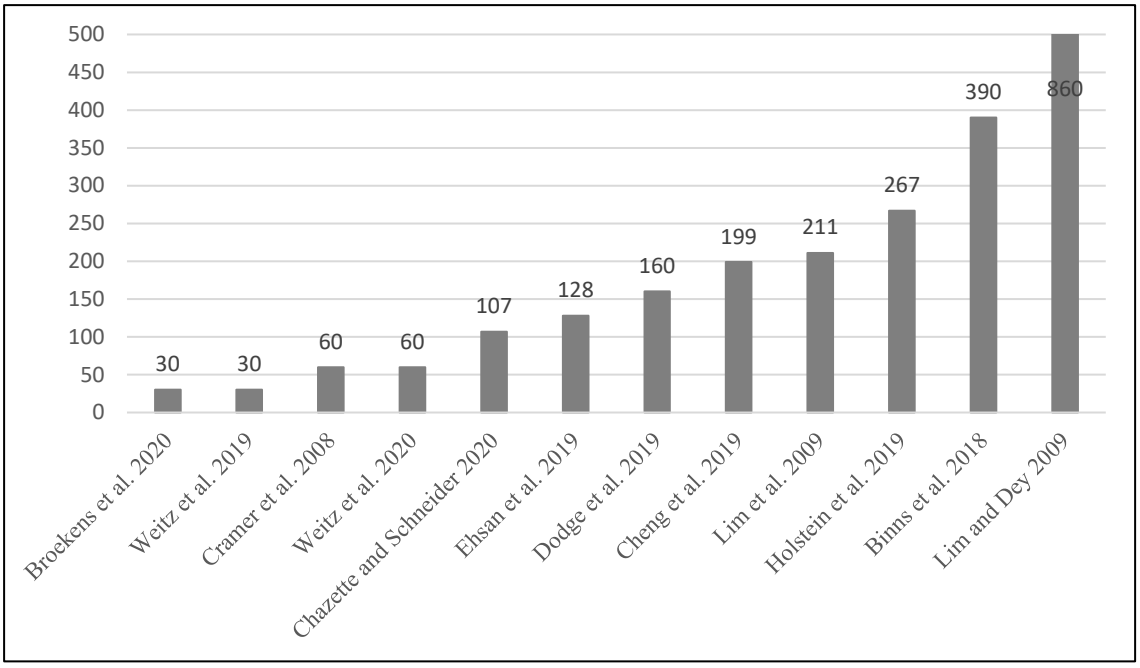

Figure 10 Sample sizes of the user survey studies

Regarding geographical distribution, United States topped the list being the country from where the empirical data was most often collected, with six studies. Austria, Brazil, India, and the United Kingdom were as well named in one study each. However, it is worth noting that in the clear majority of the articles, the country of data origin was not explicitly stated. Many studies used Amazon Mechanical Turk or another crowdsourcing marketplace to conduct their user studies, making the data collection global. Thus, it is not convenient or even possible for those studies to provide a complete list of the participants' origins.

The application areas of the 29 articles were various. The most common application area was medical diagnostics with four items. Recommender systems, context-aware applications, speed recognition, and XAI design practices of the professionals were also repeated more than once among the sample. Regarding focus groups, the end-user-themed articles lead the table with 20 items, followed by domain experts with five and developers with four papers.

Regarding the use of terms, there was a clear distinction between the articles that explicitly mention XAI and those that did not. The articles discussing XAI were generally newer than the ones that did not. One of the earliest papers to mention XAI was the article written by Broekens et al. (2010); it was the only article referring to XAI published before 2015. However, in the whole search sample, the earliest mention dated back to 2004, when proposed in the article of Gomboc et al. (2004).

Of the 29 articles, even 11 articles did not mention XAI, and nine of them were found through additional snowball sampling. Instead, the papers used keywords such as transparency (Cramer et al. 2008; Binns et al. 2018; Eiband et al. 2018; Eslami et al. 2018; Ngo et al. 2020), reliability (Bussone et al. 2015), fairness (Holstein et al. 2019), intelligibility (Lim and Dey 2009; Lim et al. 2009), trust (Yin et al. 2019) and interpretability (Eslami et al. 2018; Hong et al. 2020).

Articles were published in four different journals and 11 various conference proceedings (see Figure 11 below). The most popular publication was the Proceedings of the CHI Conference on Human Factors in Computing Systems, via which 11 articles out of the 29 in total were published. This fact was not surprising, as the conference is one of the largest and the most prestigious in the field. The second most popular venue was the Proceedings of the International Conference on Intelligent User Interfaces, with four articles. Other journals or conference proceedings had only one or two items. In total, only four articles were published in journals, whereas conference proceedings had 24 papers. This finding also reflects that the publication of articles in the field has so far focused mainly on conference proceedings.

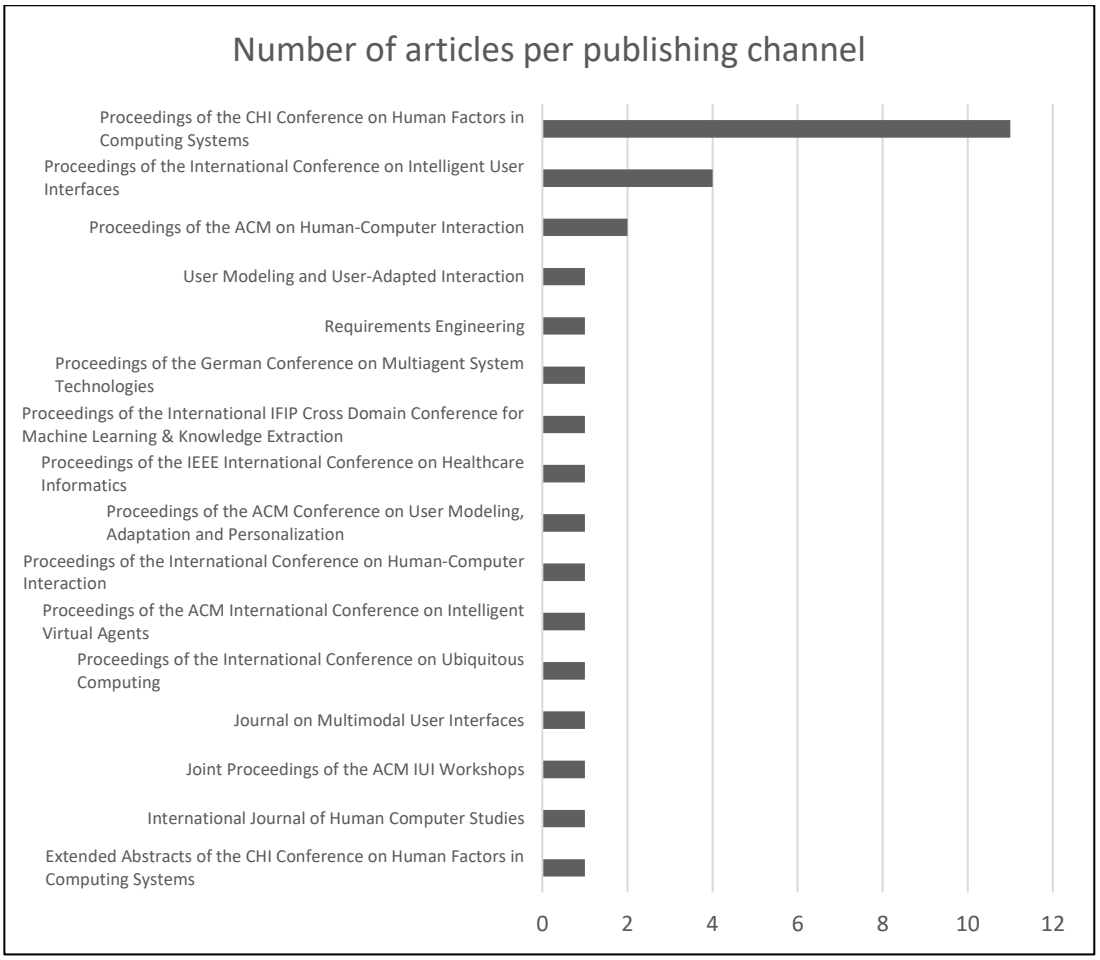

Figure 11 Publishing channels of the articles

3.2 Synthesis of the key findings based on stakeholder group

In this section, the articles' key findings are grouped and analyzed based on the articles' stakeholder group (developers, domain experts, and end-users). Of the 29 papers, four pieces focused on the developers, five on the domain experts, and twenty on the end-users.

The studies were grouped based on persons in their focus, and the groups were named after that, following the iterative process described in section 2.3. The developer-group focuses on the perspectives of the practitioners and their teams creating and developing AI solutions. Another term used for this group was data scientists. The domain experts group consisted of studies focusing on specialists in their field (another than AI), such as medical professionals. Cirqueira et al. (2020), for example, state that domain experts "are knowledgeable about their domain but not AI inner workings". The third group, end-users, consists of studies focusing on consumers, ordinary people, and other groups that use artificial intelligence solutions but are not actively involved in their design processes. For example, Weitz et al. (2019) state that their motivation to study the end-users' perspective stemmed from the fact that "much of the current XAI research is focused on machine learning practitioners and engineers while omitting the specific needs of end-users".

The findings of the three stakeholder groups were discussed on their own and further grouped. The goal was to find what kind of themes the articles discussed within each stakeholder group. The grouping began with the familiarization phase, where the findings were once again read carefully through. Based on that, a draft for thematic clusters was made for each of the stakeholder groups. The clusters were further refined and narrowed for each group in the iteration phase until the outcome was satisfactory.

An overview of the stakeholder groups and the groups' themes is provided below in Figure 12. An overwiev listing as well every article discussing particular theme is provided as Appendix 2. The findings of each stakeholder group, divided into clusters, are presented in the following sub-sections. A concluding section then discusses the similarities of the clusters between the stakeholder groups.

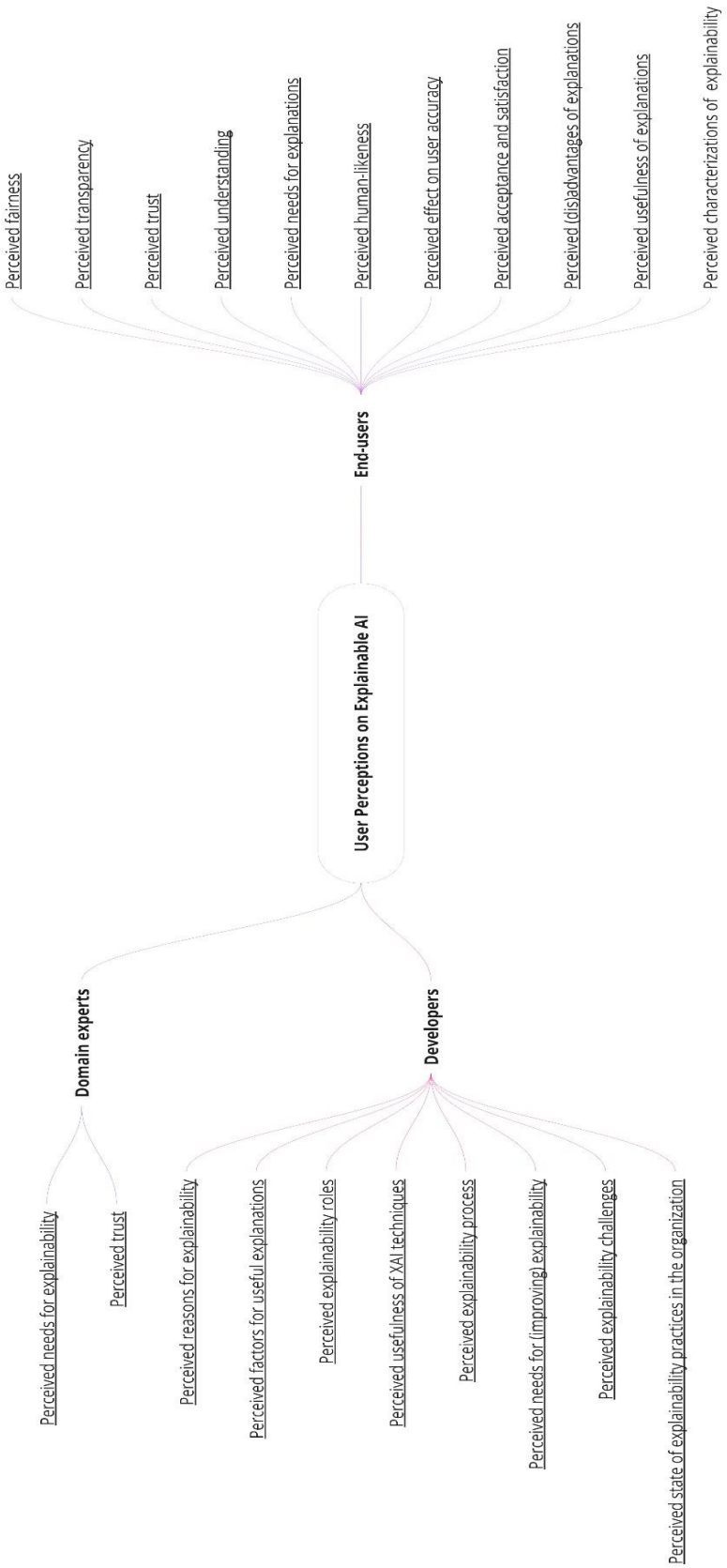

Figure 12 Overview of the stakeholder groups and themes discussed

3.2.1 Developers

A total of four articles dealt with Explainable AI from the perspective of AI, ML, or software developers. All four articles were written by scholars from the United States during the years 2019 or 2020. At the thematic level, eight different topics emerged from the papers that were discussed from the perspective of the developers. The themes were divided between articles as follows:

Discussed themes	**Articles**
Perceived reasons for explainability	• Hohman et al. 2019 • Hong et al. 2019 • Liao et al. 2020
Perceived explainability challenges	• Holstein et al. 2019 • Liao et al. 2020
Perceived needs for (improving) explainability	• Holstein et al. 2019 • Hong et al. 2019 • Liao et al. 2020
Perceived explainability process	• Holstein et al. 2019 • Hong et al. 2019
Perceived state of explainability practices in the organization	• Holstein et al. 2019
Perceived factors for useful explanations	• Liao et al. 2020
Perceived usefulness of XAI techniques	• Hohman et al. 2019

Table 2 Discussed themes within the developer group

3.2.1.1 Developers' perceived reasons for explainability

Hohman et al. (2019) find four reasons why data scientists need model interpretability: they need interpretability to generate hypotheses about the data, to understand the data, to communicate model properties to different stakeholders, and to improve the model building as well. Hong et al. (2019) also propose three broad interpretability goals that

the developers thought relevant. While they identify that interpretability is needed for decision-making, coming across new knowledge, gaining confidence, and obtaining trust, they also see that it is required for model validation and improvement.

Liao et al. (2020) looked at the same issue from a different perspective: they listed goals driving user demands for explainability from developers' standpoint. Thus, they did not map what explainability meant for the developers themselves. Their findings show that users need explainability to gain further remarks or evidence of the AI's workings, evaluate the AI system's capabilities, and adapt their behavior to better support AI utilization. The participants in their study also saw explainability as their moral requirement; providing those is their duty.

3.2.1.2 Developers' perceived explainability challenges

No common findings were found on the theme of perceived explainability challenges. Whereas the interviewees Holstein et al.'s (2019) study reported that their teams often face troubles detecting fairness issues that were not expected and fear that unexpected side effects can hinder addressing fairness issues, the participants in Liao et al.'s (2020) study see that the lack of suitable XAI tools and techniques often prevents them from satisfying the user needs. As another challenge, the interviewees in Liao et al.'s (2020) study identified that other product goals, such as legal requirements, might be at odds with their desired level of explainability. Participants in Holstein et al.'s (2019) study pondered whether those humans embedded in the model's test phases, such as crowd workers, can cause challenges to the process if the sample working on the model is biased.

3.2.1.3 Developers' perceived needs for (improving) explainability

When it comes to perceived needs for improving explainability, both Holstein et al. (2019) and Hong et al. (2019) state that developers need better tools to support them in enhancing AI explainability. The interviewees in Holstein et al.'s (2019) study state they would need tools when identifying which sub-populations they should consider in model creating to ensure that the dataset will become balanced. The participants in their study also complained about the limitations of the auditing processes – more holistic, system-level methods would be needed in their opinion and domain-specific auditing. The participants in Hong et al.'s (2019) study would like to have tools to help them understand how a particular model makes decisions. These tools would be needed particularly in three situations: in conducting root cause analysis, recognizing decision boundaries, and identifying the global structure so that the model's operation can be described.

Regarding tools, Holstein et al. (2019) also propose that the practitioners would like to have resources that would enable them to learn from others' experiences, given the limited amount of time and resources they have. Tools that would pool fairness knowledge between different teams were found out to be very useful.

In their part, Liao et al. (2020) state that developers would need guidance in explainability needs specification. They would also like to receive support in creating explainability solutions, coupled with example artifacts, which would support the communication between stakeholders.

3.2.1.4 Developers' perceived explainability process

Two articles, Holstein et al. (2019) and Hong et al. (2019) discussed the perceived explainability process. The discussions of the items are not proportional. Holstein et al. (2019) merely mention the process; they found out that most interviewees look first to their training datasets before the models when they try to improve their product's fairness. On the other hand, the study of Hong et al. (2019) maps the process meticulously. They identify three interpretability roles and the same amount of interpretability stages based on the interviews. The three roles identified are model builders, model breakers (domain specialists), and model consumers. The three interpretability stages Hong et al. (2019) identify are the ideation and conceptualization stage, building and validation stage, and deployment, maintenance, and use stage. The stages represent different phases of the actual product development process. The authors also find four themes that characterize interpretability work throughout the product's life: it is a process, it is co-operative, it is context-dependent, and it is about comparing mental models.

3.2.1.5 Other themes discussed

Three themes were discussed only by one article. Holstein et al. (2019) examined the perceived state of the explainability practices in the organization. 65 % of the respondents of their survey stated that their teams can have at least some control over the data collection process and that 58 % currently consider fairness at the stages of data collection. However, most interviewees stated that they do not have functions that would support them in collecting balanced datasets. Interestingly, interviewees also noted that they are not rewarded for their efforts for improving fairness, and only 21 % of the survey respondents stated that their teams currently prioritize fairness to a great extent.

Liao et al. (2020) discussed perceived factors for valuable explanations. They identified few factors the developers considered relevant for desirable explanation:

explanations should be selected, focusing only on few factors, and that explanations are inherently social. The authors imply that these two characterizations – selective and social – mean that XAI should be seen as interactive or conversational. Lastly, Hohman et al. (2019) discussed the perceived usefulness of particular XAI techniques. When it comes to global and local explanations, the study finds them complementary. Data scientists' preferences about explanation strategies correlate with their expertise: the more novice developers preferred local explanations. In contrast, more senior colleagues used global ones more frequently. The most senior participants in the study used both.

3.2.1.6 Synthesis of the developer cluster

The heterogeneity of the four studies means that it is challenging to find overlapping themes between them. The study of Hohman et al. (2019) had a different perspective than the others since it focused on what interpretability meant to the developers themselves. In contrast, the other studies were more focused on the user's viewpoint and the process in which the practitioners tried to provide explainability to them. Then again, the study of Holstein et al. (2019) is very focused on the developers' troubles and needs when improving explainability. In contrast, Hong et al. (2019) deepened into the process in which explainability is created. Liao et al. (2020) focused on narrowing the rift between XAI research and practical AI design work.

Of the two studies investigating perceived explainability challenges the practitioners face, there were no overlapping findings. Liao et al. (2020) focused on the challenges around providing adequate answers to user questions and the lack of suitable tools. In contrast, Holstein et al. (2019) pondered the challenges unexpected effects could cause. Holstein et al. (2019) then focused on showing the most influential stages for fairness interventions for the perceived explainability process. In contrast, Hong et al. (2019) investigated and specified the actual steps that formed the process.

However, there are few overlapping observations in the emerged themes that were shared between articles. As perceived reasons for explainability, both Hong et al. (2019) and Liao et al. (2020) constate that in the eyes of the developers, users need explainability to make more informed decisions and to get a grasp of the capabilities of the system.

When it comes to perceived needs for improving explainability, all three articles (Holstein et al. 2019; Hong et al. 2019; Liao et al. 2020) hint that the practitioners would need better tools to support their quest of providing explainability. However, the domains of the tools differ: whereas Holstein et al. (2019) propose that developers need tools for providing a balance between the sub-populations in datasets and better auditing methods, Liao et al. (2020) constate that practitioners need guidance in explainability needs

specification and Hong et al. (2019) see that practitioners themselves need tools to understand how a particular model makes decisions.

Although no significant similarities were found between the articles' findings, these four studies provide a glimpse into how the developers see the field of XAI from their perspective. As for future research avenues, these four studies provide a good and exciting starting point for studies exploring how explainability is created, its challenges, and the practitioners' needs through the process.

3.2.2 Domain experts

Five articles explored XAI from the perspective of domain experts. Four of the studies explored XAI from medical professionals' perspective (Bussone et al. 2015; Cai et al. 2019; Wang et al. 2019; Xie et al. 2019), whereas one (Cirqueira et al. 2020) studied XAI through the lens of fraud detection specialists in banking. All studies were qualitative and used either interviews or user experiments as their research method and were written during 2019 and 2020 except for Bussone et al.'s (2015) article. Two major explainability themes emerged from the papers. The following table demonstrates the distribution of themes between articles.

Discussed themes	**Articles**
Perceived needs for explainability	• Cai et al. 2019 • Cirquiera et al. 2020 • Wang et al. 2019 • Xie et al. 2019
Perceived trust	• Bussone et al. 2015 • Wang et al. 2019 • Xie et al. 2019

Table 3 Discussed themes within the domain expert group

3.2.2.1 Domain experts' perceived needs for explainability

All articles except Bussone et al. (2015) discussed the domain experts' perceived needs for explainability. Cai et al. (2019) identified a list of five themes around which the needs of pathologists centered: capabilities and limitations of the diagnostic tool (e.g., edge cases, diversity of training data), its functionality (e.g., algorithm's access to context, the steps of its analysis of inputs), its medical point-of-view (e.g., how the algorithm would

judge borderline cases, its source of medical ground truth), its design objectives (e.g., whether it is meant for supplementing human or working alone), and the considerations needed before adapting the tool (e.g., regulatory approval, cost of purchase).

Xie et al. (2019) identify three critical steps to interviewees' interaction with the data: detecting the borderline cases (like Cai et al. 2019), generating prioritization matrices, and coordinating with (other) computer-based systems. Wang et al. (2019), for their part, propose six design suggestions to (medical) XAI systems, both to answer to user needs and increase their trust: user's hypothesis generation must be supported, forward, data-driven reasoning must be supported, coherent factors must be supported, access to source and situational data must be supported, Bayesian reasoning must be supported, and multiple explanations should be integrated into single explanations.

Cirqueira et al. (2020) found out that, from the perspective of fraud detection specialists, the explanations should be selective, combining only the most essential pieces of evidence, processing the explanations should not require a heavy mental workload, and explanations should enable them to perform local comparisons of fraud cases.

3.2.2.2 Domain experts' perceived trust

Three articles discussed the domain experts' perceived trust towards XAI. Bussone et al. (2015) found out that confidence explanations (how sure the system is that its suggestion is correct) did not increase the participants' trust towards the system but can still help the medical professionals assess the system's reliability. Whereas comprehensive why explanations (why the system made the suggestion) did increase trust towards the system but promoted even over-reliance, selective why explanations, on the other hand, promoted self-sufficiency, which can both result in potentially wrong diagnoses.

Xie et al. (2019) propose that to foster trust, medical professionals should be allowed to manipulate the user interface to choose and prioritize between different types and sources of data. They state that medical AI systems should be gradually incorporated into the diagnosis process. Wang et al. (2019) also state that the domain experts should have access to source and situational data, as discussed earlier.

3.2.2.3 Synthesis of the domain expert cluster

Since Cirqueira et al.'s (2020) article was the only one discussing fraud detection specialists' perspective, its observations cannot be generalized to a broader discussion. More common ground can be found between the four articles discussing medical professionals' perspectives. No contradictory findings emerged from the articles discussing medical

professionals' viewpoints. When it comes to factors affecting clinicians' perceived trust towards XAI systems in clinical diagnostics, both Wang et al. (2019) and Xie et al. (2019) propose that to foster trust in the system, the users should have access to source data. Xie et al. (2019) even propose that clinicians should be able to prioritize between different sources and types of data. Bussone et al. (2015) propose that users' trust towards the system would increase if shown differential diagnoses, confirmed by Wang et al.'s (2019) viewpoint that users' hypothesis generation should be supported by giving them contrastive explanations.

In terms of perceived user needs for explainability, access to the background data is also highlighted. Whereas Wang et al. (2019) and Xie et al. (2019) emphasize the access to the data and the possibility to choose between data sources, Cai et al. (2019) state that the users need at least to know how diverse the training data was. Whereas Cai et al. (2019) propose that users want to have multiple explanations to support their diagnostic process, Wang et al. (2019) state that users' forward, data-driven reasoning should be supported, running "counter to most recommendations of showing shallow explanations first".

Based on these findings, it is safe to say that medical professionals tend to prefer receiving rather too much information than too little. This finding is in stark contrast with the Cirquiera et al.'s (2020) finding that fraud detection specialists want to have just the right amount of selective information. Medical professionals prefer meticulousness in their decision-making, while speed is the primary quality criterion in decision-making for fraud detection professionals. This finding highlights how the needs for explainability vary between different domains, depending on the unique needs of each domain expert group.

3.2.3 End-users

In total, nineteen articles explored end-user perspectives on XAI. The findings of the articles could be grouped under nine different themes. The distribution of the articles based on the themes discussed is depicted in the following table.

Discussed themes	**Articles**
Perceived fairness	• Binns et al. 2020 • Dodge et al. 2019
Perceived trust	• Cheng et al. 2019 • Lim et al. 2009 • Schrills and Franke 2020 • Weitz et al. 2019

	• Weitz et al. 2020 • Yin et al. 2019
Perceived understanding	• Cheng et al. 2019 • Cramer et al. 2008 • Ehsan et al. 2019 • Eiband et al. 2018 • Eslami et al. 2018 • Lim and Dey 2009 • Lim et al. 2009 • Ngo et al. 2020 • Oh et al. 2018
Perceived human-likeness	• Ehsan et al. 2019 • Oh et al. 2018
Perceived acceptance and satisfaction	• Ehsan et al. 2019 • Eslami et al. 2018 • Lim and Dey 2009 • Putnam and Conati 2019 • van der Waa et al. 2020
Perceived transparency	• Ngo et al. 2020 • Schrills and Franke 2020
Perceived needs for explainability	• Brennen 2020 • Chazette and Schneider 2020 • Eiband et al. 2018 • Eslami et al. 2018 • Lim and Dey 2009 • Oh et al. 2018 • Putnam and Conati 2019 • van der Waa et al. 2020 • Weitz et al. 2020
Perceived usefulness of explanations	• Broekens et al. 2010 • Lim et al. 2009
Perceived advantages and disadvantages of explanations	• Chazette and Schneider 2020
Perceived characterizations of explainability	• Brennen 2020

Table 4 Discussed themes within the end-user group

The particular case among the end-user papers was the article of Brennen (2020). It explores how different stakeholders in AI and ML industries characterize Explainable AI and what they want from it. This article was first categorized under its own group, other stakeholders, but later it was decided to combine with the end-user cluster. This move was made because one article was not seen enough to combine its own category and because there were end-users among the interviewees of the article. The author conducted 40 interviews, participants of which consisted of company founders, investors, potential end-users, and academia members, for example.

3.2.3.1 End-users' perceived fairness

Two articles discussed the end-users' conceptions around the perceived fairness of AI. Although the discussions around fairness were closely related to perceived trust and transparency, these were considered separate categories since the terms have a slightly different emphasis. Both Binns et al. (2020) and Dodge et al. (2019) make two similar findings. In both studies, case-based explanations were seen generally as less fair than other explanation styles, whereas sensitivity-based styles were the most effective ones. Case-based explanations use the knowledge accumulated from previous occurrences, in other words, cases, and choose the most suitable one from the model's training data to describe the situation at hand (Binns et al. 2020). On the other hand, sensitivity-based styles show how much some or all the input variables in the model would need to change to change the output class – in other words – the decision to some other (Binns et al. 2020).

Both studies emphasize the end-users' prior conceptions about AI as highly determinative to their judgment of fairness. Binns et al. (2020) noted that the very idea of algorithmic decision-making seemed to be unfair to some participants of their study. In contrast, Dodge et al. (2019) constate that an individual's prior conceptions have a "significant impact on how they react to explanations, and possibly more so than differences in cognitive styles".

3.2.3.2 End-users' perceived trust

Altogether six articles discussed factors that affect end-users' perceived trust towards XAI systems. Two of the studies were written by the same research group (Weitz et al. 2019 & 2020) and discussed virtual agents' potential in increasing users' trust in AI systems. In contrast, other articles studied trustworthiness from the perspectives of explanation interfaces in university student admissions (Cheng et al. 2019), context-aware

systems (Lim et al. 2009), visual explanations (Schrills and Franke 2020), and prediction tasks (Yin et al. 2019).

Both articles from Weitz et al. (2019 & 2020) give the same conclusions about using virtual agents in explanations. In both studies, end-users had more trust in explanations shown by virtual agents than, for example, to only text- or voice-based explanations. However, as the same research group conducted both studies in Germany, and no other article in the sample studied the effect of virtual agents, further work and confirmation on the findings is still needed.

An interesting finding from the study of Cheng et al. (2019) is that the explanation interfaces enhanced users' understanding of the model but not its trustworthiness. The authors ponder that this could be because of users' discomfort in using algorithms to make university admissions decisions. This finding supports Binns et al. (2020) and Dodge et al. (2019) findings of the importance of users' prior conceptions of AI to the perceived fairness. Users will not trust the model or consider it fair regardless of the improvements made to it if they consider the system's task type fundamentally unfit for algorithmic decision-making.

A further finding to the same theme from the study of Yin et al. (2019) is that users' trust towards the AI did not increase even in the situation where the system's observed accuracy was higher than the participant's own accuracy if the system's stated accuracy was substantially higher than its observed one. This result also reflects users' per se suspicion towards AI. AI systems are thus considered trustworthy if they match users' conceptions about what sectors of life are suitable for algorithmic decision making and if those systems offer very reliable information about their accuracy.

Both Schrills and Franke (2020) and Lim et al. (2009) study the effects of different explanation types on users' perceived trust. Whereas counterfactual explanations that Schrills and Franke studied are closely related to "what if" -explanations ("what would system do if X") Lim et al. studied, those two types are not strictly comparable. Thus, summaries between observations of the two studies cannot be made. But when looking over the focus group boundaries, a common finding between Lim et al.'s (2009) study and Bussone et al.'s (2015) study about domain experts is that, in both studies, why explanations (why the system made the particular suggestion) are the most compelling explanation styles in fostering users' trust towards AI systems.

3.2.3.3 End-users perceived understanding

Almost half of the articles, nine in total, discussed the end-users perceived understanding of AI systems. The application fields of the studies varied from fitness and gaming applications (Eiband et al. 2018; Ehsan et al. 2019) to online advertising (Eslami et al. 2018),

university admissions (Cheng et al. 2019), and context-aware applications (Lim and Dey 2009; Lim et al. 2009). Two studies focused on recommender systems on art and movies (Cramer et al. 2008; Ngo et al. 2020, respectively), and one on human-AI art co-creation (Oh et al. 2018). Except for one study, the articles' findings centered on which factors or explanation types affected the users' perceived understandability towards XAI systems. In contrast, one study (Ngo et al. 2020) studied users' mental models and how the process towards understanding is formed in their minds.

Many findings focused on the effects of language and context of the explanations on users' understanding. For example, Eslami et al. (2018) found out that obscure and over-simplified language hindered participants' understanding of explanations, whereas Oh et al. (2018) state that users preferred detailed yet selective information over basic information or too many instructions. Ehsan et al. (2019) also state that it is the contextual accuracy rather than the length of the explanation which matters most to the users' understanding. They even propose that the more relatable the rationale is, the better it is for users understanding. Explanations must thus be formed so that the language and the context are relatable to the users and that they do not overwhelm users. The relatability is also in line with the findings of Ngo et al. (2020) regarding the recommender systems: the authors formulate a concept of "centrality of self"; end-users want that the recommender system experience is centered around themselves. The other studies list, for example, interactivity (Cheng et al. 2019) and transparency (Cramer et al. 2008) as factors that increased users' perceived understanding.

3.2.3.4 End-users' perceived human-likeness

Two of the studies mentioned above, Ehsan et al. (2019) and Oh et al. (2018), discussed the explanations and AI's perceived human-likeness. In these studies, human-likeness was considered as a positive, desirable characteristic to AI systems. Oh et al. (2018) found out that their study participants tended to anthropomorphize the AI and considered it as a genuine personality, although sub-ordinate to humans. The interviewees, however, did not like when the AI switched between human- and machine-like; they wanted it to be relatable throughout the process. When it comes to relatability, Ehsan et al. (2019) state that participants found out explanations more human-like when those "mirrored their thoughts", which is in line with the concept of "centrality of self" discussed in the previous paragraph.

3.2.3.5 End-users' perceived acceptance and satisfaction

Five articles discussed themes of user acceptance and satisfaction. Besides three already discussed articles by Ehsan et al. (2019), Eslami et al. (2018), and Lim and Dey (2009), the two other articles discussed tutoring systems (Putnam and Conati 2019) and decision support systems in autonomous cars (van der Waa et al. 2020). Starting with Ehsan et al. (2019), one interesting observation they discussed in their study was that the more detailed explanations inspired more confidence, being more intelligent and predictive. This observation might reflect gamers' needs more broadly: they are ready to invest their time and money into gaming and want to have everything out of the experience; thus, the need for detailed explanations. Combining this finding with Oh et al.'s (2018) previous observations about understandability – participants understood explanations best when those were detailed yet selective – it seems that end-users in a creative context, such as art creation and gaming, prefer detailed explanations. However, as these findings were based only on two studies, more research in the field of the complexity of explanations is needed to understand its effects fully.

Regarding explanation styles and needs, participants in van der Waa et al.'s (2020) study were most satisfied when the explanations were made based on generalized prior experiences rather than showing them just one particular prior case. Eslami et al. (2018) found out that participants were most satisfied with explanations that showed particularly why the ad was targeted to them. In contrast, participants in Putnam and Conati's (2019) study found explanations most valuable when they were incurious mood themselves. Since the studies and their findings were very homogenous, drawing conclusions about user satisfaction factors would require more research.

3.2.3.6 End-users' perceived transparency

Only two articles discussed perceived transparency, although the themes discussed were closely related to fairness and trust discussions. Schrills and Franke (2020) brought out that visual explanations increased the stated transparency. In contrast, Ngo et al. (2020) found out that Netflix's recommender system was not considered very transparent and even led the participants to a "gulf of execution", left unsure what to do next. Since the research field is still very new, it is not surprising that the number of transparency-related studies was meager. However, it is worth noting that transparency was mentioned in several other studies in the field, but their contributions did not focus on transparency.

3.2.3.7 End-users' perceived needs for explainability

Altogether nine articles discussed the role of users' perceived needs for explainability. The only article not discussed before was the one written by Chazette and Schneider (2020), exploring explainability as a non-functional requirement in the context of navigation systems. As much as three articles discuss the importance of why explanations (why the system did action X) to the end-users. Putnam and Conati (2019) state that participants in their study wanted why explanations most often, and Chazette and Schneider also discovered the same result. These findings are in line with the conclusion of Lim and Dey (2009) that some "intelligibility" (or explanation) types, such as why should be incorporated into all context-aware applications. Another common finding was that the end-users were supportive of the explanations; they stated that explanations were needed (Chazette and Schneider 2020; Eiband et al. 2018; Oh et al. 2018; Putnam and Conati 2019; Weitz et al. 2020).

Because the studies discussing end-users' needs for explainability were very heterogenous, there were no further shared findings. However, there were some additional interesting findings in the individual studies, such as that end-users want that explanations are fit for for intuitive comparisons (Weitz et al. 2020) or that explanations should be presented only when the user demands them (Chazette and Schneider 2020), or that end-users need explanations especially in situations where they do not agree with the system's perceived decision making or decision-making process (Putnam and Conati 2019).

One interesting finding came from the article of Brennen (2020), which discussed these themes from a slightly different perspective. The author proposes several different "use cases" for XAI, of which she identifies three: debugging models, identifying bias, and building trust. The author also suggests that the stakeholders want more transparency into complex ML models' operation, more knowledge about biases that might be included in the training data, model, or the deployed system. They also need explainability to foster their trust in technologies that unfamiliar to them. The author proposes that future XAI tools need to take all the different use cases into account since different audiences need different explanations.

3.2.3.8 End-users' perceived usefulness of explanations

Two articles – Lim et al. (2009) and Broekens et al. (2010) – discussed users perceived thoughts about the usefulness of the explanations they received. The article of Broekens et al. discussed the usefulness and naturalness of explanations in the context of autonomous (cooking) systems, whereas Lim et al. (2009) studied different "intelligibility" types. Broekens et al. (2010) found out in their study that more experienced cooks judged

the explanations more valuable than more novice, which they highlight for its counter-intuitiveness. They also state that explanations are helpful if they are tailored to different situations. Lim et al. (2009), on their behalf, found out that how to and what if explanations might be helpful in situations where the end-user tries to figure out how to execute some system functionalities. There were no shared observations between the studies that could have been made.

3.2.3.9 Other themes discussed

The last two categories, perceived advantages and disadvantages of explanations, and perceived characterizations of explainability were only discussed by one article each. As perceived advantages of explanations, Chazette and Schneider (2020) found that explanations can facilitate understanding of the system, clarify doubts and reduce obscurity, and provide decision-making support. In contrast, as disadvantages, they state explanations can hinder understanding, provide unnecessary information and add more obscurity. These are as well in line with many findings from the needs and understanding sectors.

Under perceived characterizations of explainability was the article of Brennen (2020). The author states that the scientific XAI discourse misses a common, shared terminology as its key finding. The participants used a broad range of synonyms for explainability and had various subjective perceptions about their meaning. Some interviewees made clear distinctions between two terms, whereas others would use those interchangeably to mean explainability. A group of interviewees included almost any technical solution handling data automatically, such as Microsoft Excel, under the umbrella of Explainable AI. These misunderstandings might well be the case why the field of XAI is, at the moment, very heterogenous and it still misses established terminology. This fact is logical since the connotations of "fair", "just", and "transparent", for example, mean different things to different people.

However, as these two studies were the only ones discussing their respective fields, comparative observations cannot be made. However, Brennen's (2020) findings are very much in line with this study's observations that XAI has started to gain a foothold in the debate in the last few years.

3.2.3.10 Synthesis of the end-user cluster

The end-user cluster was highly heterogeneous, and its application areas various. However, some interesting, more general findings could be made. Based on this analysis, it can be said that case-based explanations were generally seen as less fair than other

explanation styles. In contrast, sensitivity-based styles were the most effective ones. A couple of studies (Dodge et al. 2019; Binns et al. 2020) also emphasized the end-users' prior conceptions about AI as highly determinative to their judgment of fairness. An interesting finding from the study of Cheng et al. (2019) is that the explanation interfaces augmented users' understanding of the model but did not increase its trustworthiness. This finding would seem to imply that increasing end-users' trust towards AI might be more complex than thought.

Another common finding was that explanations must be formed so that the language and the context are relatable to the users and that they do not overwhelm users. End-users also seem to understand explanations best when those were detailed yet selective – at least in creative contexts. It does not come as a surprise either that the end-users were supportive of the explanations, and they stated that explanations were needed. Explanations are thus worth pursuing.

Some interesting findings from individual studies that would be worthy of more research were that end-users want explanations that are befitting for intuitive comparisons (Weitz et al. 2020), or that explanations should be presented only when the user demands it (Chazette and Schneider 2020) or that end-users need explanations especially in situations where they disagree with the AI's perceived decision making or decision-making process (Putnam and Conati 2019). Since these findings were only confirmed by one study each, more research on these would be needed.

3.3 Addressed research questions and theoretical approaches

3.3.1 Classification of research questions

The studies' research questions were categorized under six different groups based on the theory of six "elementary question forms" formulated by Bunge (1967). The six question categories are which, where, why, whether, how, and what. The most popular question categories within the sample of this study were "whether" and "how", whereas no research questions were categorized under the classes of "where" and "why". The absence of why questions may reflect current XAI research's pragmatism: often, articles focused on a particular technical solution, its effectiveness, and modification of its various features. Additionally, one article's status was categorized as "unclear" since it was challenging to deduct the article's actual research question. The research questions of the studies divided into their respective categories are depicted in the following table 5.

Research question type	Article	Research questions
Which	Chazette and Schneider 2020	"RQ2: Which nonfunctional requirements related to transparency can be impacted by the problems described by the end-users?"
	Lim and Dey 2009	"which types of questions users want answered [when using context-aware intelligent system]"
	Weitz et al. 2020	"2. Which of the three modalities of a virtual agent that we tested (pure information in form of text, voice, and visual presence) are important for an impact on the perceived trustworthiness of an AI system?"
Where	–	–
Why	–	–
Whether	Broekens et al. 2010	" [Whether] different explanation types are needed to explain different types of actions."
	Chazette and Schneider 2020	"RQ4: Are there significant differences between the opinions of digital immigrants and natives regarding explanations in software systems?"
	Cheng et al. 2019	"Research Question 4: Will explanation interfaces increase users' trust in the profiling algorithms?"
	Cramer et al. 2008	"whether transparency increases users' trust and acceptance of recommender systems."
	Dodge et al. 2019	"RQ1a Are some explanations judged to be fairer? RQ1b Are some explanations more effective in surfacing unfairness in the model? RQ1c Are some explanations more effective in surfacing fairness discrepancies in different cases?"
	Ehsan et al. 2019	"[whether] human-like plausible rationales can be generated using a non-synthetic, natural language corpus of human-produced explanations."
	Oh et al. 2018	"[Whether] users [would] like to take the initiative or let AI take it when they cooperate [in creative contexts]?

	Putnam and Conati 2019	"—[whether] if it is necessary for an Intelligent Tutoring System (ITS) to explain its underlying user modeling techniques to students."
	Weitz et al. 2019	"Does the incorporation of a virtual agent into XAI approaches positively impact the perceived trustworthiness of complex intelligent systems like Deep Neural Networks (DNN)?"
	Weitz et al. 2020	"1. Does the usage of a virtual agent positively impact the perceived trustworthiness of AI systems like deep neural networks?"
	Yin et al. 2019	"Does a model's stated accuracy on held-out data affect people's trust in the model? If so, does it continue to do so after people have observed the model's accuracy in practice?"
How	Binns et al. 2018	"1. How do explanations for algorithmic decisions affect justice perceptions regarding algorithmic decisions? In particular, do the positive correlations observed between informational, procedural and distributive justice in human decision-making settings also hold in algorithmic decision-making settings? 2. How do different styles of explanation affect such justice perceptions?"
	Cheng et al. 2019	"Research Question 1: How effective are the white-box and black-box strategies in helping non-expert users understand profiling algorithms? Research Question 2: How effective are the interactive and static interfaces in helping non-expert users understand profiling algorithms? Research Question 3: How will users' personal characteristics (i.e. education level and technical literacy) influence the effectiveness of the explanation interface in helping them understand profiling algorithms?"
	Dodge et al. 2019	"RQ1 How do different styles of explanation impact fairness judgment of a ML system? RQ2 How do individual factors in cognitive style and prior position on algorithmic fairness

		impact the fairness judgment with regard to different explanations?"
	Eslami et al. 2018	"RQ1: a) How do users perceive and evaluate existing ad explanations? b) Given the opportunity to craft their own ad explanations, how do users' preferred ad explanations compare to the existing ad explanations? RQ2: When exposed to typically hidden inner attributes of an algorithmic advertising platform (such as users' algorithmically-derived attributes and how advertisers use them), how do users think about and evaluate these attributes?"
	Holstein et al. 2019	"How might educational AI (AIED) systems best be designed to support these complementary roles [of students and teachers]?"
	Lim and Dey 2009	"–how answering them [questions users want to be answered] improves user satisfaction of context-aware applications."
	Oh et al. 2018	"How do users and AI communicate in creative contexts?"
	Schrills and Franke 2020	"How different prototypical visualizations that aim to explain AI results affect the perceived trustworthiness and observability of an image classification system?"
	Weitz et al. 2020	"3. How are the presented XAI visualisations perceived and rated by users? 4. How does the use of virtual agents affect the perception of the presented XAI visualisations?"
	Xie et al. 2019	"RQ: How do medical professionals interact with patients' data for diagnosis and/or treatment purposes?"
	Yin et al. 2019	"How does a model's observed accuracy in practice affect people's trust in the model?"
What	Chazette and Schneider 2020	"RQ1: What is the current perception of end-users regarding the need for explanations? RQ3: What are the advantages and disadvantages of receiving explanations and how do they relate to transparency?"

	Dodge et al. 2019	“RQ3 What are the benefits and drawbacks of different explanations in supporting fairness judgment of ML systems?”
	Holstein et al. 2019	"As artificial intelligence (AI) increasingly enters K-12 classrooms, what do teachers and students see as the roles of human versus AI instruction?”
	Ngo et al. 2020	“RQ1: What are the mental models users hold of a RS (recommender systems)? RQ2: To what extent is the RS perceived as transparent? RQ3: To what extent is the RS perceived as controllable? RQ4: What implications for RS design can be derived?”
	Oh et al. 2018	“What factors are associated with the various experiences in this process [of users and AI collaborating in creative contexts]?"
Unclear	Eiband et al. 2018	"Our aim is to advance existing UI guidelines for more transparency in complex real-world design scenarios involving multiple stakeholders. To this end, we contribute a stage-based participatory process for designing transparent interfaces incorporating perspectives of users, designers, and providers, which we developed and validated with a commercial intelligent fitness coach."

Table 5 Overview of the research questions of the studies

3.3.1.1 Which questions

Three studies approached their research subject through the “which” question. The application areas and the objectives for the question differed widely. Chazette and Schneider (2020) tried to find out in their second research question (of the total four), which non-functional requirements (in the system architecture) connected to transparency are affected by the difficulties reported by the end-users (of navigation systems). Lim and Dey (2009) searched answers to which questions users want to be answered when using context-aware systems. In contrast, Weitz et al. (2020) asked in their second research

question (of the total three), which of the three modalities (text, voice, virtual presence) they tested affect the perceived trustworthiness of their AI system (explained by virtual agent).

Thus, two studies explored transparency-related factors. Chazette and Schneider (2020) found out that the problems users described were related to non-functional requirements' usability and informativeness. Weitz et al. (2020) on their part state that users' trust was increased by speech output when compared to text output and even further increased when a virtual agent presented the results rather than raw speech output. In the third study, Lim and Dey (2009) found out that users especially want *why* questions (why something happened) answered and need *certainty* information (how confident a prediction is). The last observation – the users' need for certainty and correctness – is, perhaps a bit surprisingly, the only common finding that can be made based on the articles, since correctness is one of the sub-dependencies to informativeness NFRs, described by Chazette and Schneider (2020).

3.3.1.2 Whether questions

Whether-related questions were the joint largest category of the question types, with in total 11 articles having at least one whether-question. In this section, the most prominent question cluster was related to users' trust-related themes. The transparency-related whether questions varied from whether explanation interfaces will increase users' trust (Cheng et al. 2019) to whether virtual agents increase users' trust (Weitz et al. 2019 & 2020) and whether transparency increases users' trust in recommender systems (Cramer et al. 2008). The fourth and the last trust-related question asked whether the model's described precision on held-out data impacts users' trust in the model (Yin et al. 2019). The explanation interfaces were found out to increase users' understanding but not their trust (Cheng et al. 2019, 559), as well as transparency itself neither did foster users' trust towards systems (Cramer et al. 2008), whereas virtual agents, however, did increase users' trust (Weitz et al. 2019; Weitz et al. 2020). In the last study, Yin et al. (2019) found out that the stated accuracy on held-out data (dataset that is split to training and testing sets) influences users' trust towards the model. Still, the effect is minor after people have observed it in action. These findings would seem to imply that the effects of current explainability solutions to users' trust towards AI systems are, in fact, relatively small or even non-existing. However, the sample here is tiny, so more research on these issues would be needed to confirm the effects.

Another question cluster was related to the effectiveness of explanations. Broekens et al. (2010) asked whether distinct explanation types are needed to explain different phenomena. In contrast, Dodge et al. (2019) asked whether some explanations bring more

effectively forth unfairness in the model and between different cases. Broekens et al. (2010) found out that their hypothesis about the need for different explanations in different situations was supported and that an action is always preferred explained by its parent goal. Determined by the action type, some supplementary information could also be needed to support the explanation. On their part, Dodge et al. (2019) found out that local explanations are better in highlighting the case-specific fairness problems or issues between cases than global explanations. What can be said based on these observations is that different explanations thus have different qualities and could be more helpful or perform better, depending on the situation.

Apart from the clusters, several research questions in the whether-category could not be grouped with others. One study (Chazette and Schneider 2020) asked whether there are differences between digital natives' and digital immigrants' opinions on software systems explanations. Their findings concluded that no statistically significant differences between the groups were found. However, the authors noted that the participants, regardless of age group, had a high technical knowledge, which might have affected the results. Another study (Oh et al. 2018) asked whether users would like to take the rein or let the AI take it in creative contexts. Their study showed that users wanted to take the lead but let the AI do independently those tasks considered dull, such as coloring.

Two more whether-type questions were found among the sample. Whereas Putnam and Conati (2019) asked whether it is mandatory for an Intelligent Tutoring System to describe its background user modeling procedures to students, Ehsan et al. (2019) studied whether human-like rationales could be created by "using a non-synthetic, natural language corpus of human-produced explanations". However, because the approaches of the articles varied so much, no unifying findings were found.

3.3.1.3 How questions

How-questions had as well 11 articles, sharing the title of the largest category with whether-related questions. The application areas and approaches chosen were as well very heterogenous in this category. In this category, one question cluster was related to justice perceptions; how explanations for algorithmic decisions affect peoples' perceptions, and what effect do explanation styles have? Two studies contributed to this discussion: Binns et al. (2018) and Dodge et al. (2019). In both studies, case-based explanations were less fair than, for example, global and sensitivity-based explanations. However, Binns et al. (2018) state that explanation styles do not usually have a notable impact on justice perceptions. They also state that the justice interrelationships from human decision-making were found out to hold with algorithmic decisions. Thus, the explanations were judged as if another human had made them. Dodge et al. (2019) also studied how users' prior

conceptions about ML affect how they react to explanations. The impact is even more significant than the differences in cognitive styles.

Another question cluster formed around XAI visualizations. Schrills and Franke (2020) asked how visualizations affect users' perceived trust in image classification systems. In contrast, Weitz et al. (2020) studied how users perceive visualizations and how virtual agents affect the perceptions of visualizations presented. Schrills and Franke (2020) found that their hypothesis about visualizations contributing to more vital trust could not be confirmed. Still, they found substantial connections between acceptances to classification and perceived trust. Weitz et al. (2020), on their part, found that visualizations that virtual agents presented were the most trustworthy in the eyes of users. Thus, visualizations and especially virtual agents, would positively affect users' trust, but more confirmatory research on these themes would be needed.

No other question clusters could be made within this category since the questions were that heterogeneous. Whereas Cheng et al. (2019) studied how effective black-box and white-box methods and interactive and static interfaces helped end-users understand profiling algorithms, asked Eslami et al. (2018) how users evaluated ad explanations. Holstein et al. (2019) then discussed how educational AI could support students and teachers in the classroom. Oh et al. (2018) studied how users and AI communicate in creative contexts, and Lim and Dey (2009) studied how answering the questions users want to be answered affects their satisfaction. Yin et al. (2019) studied what effect the model's accuracy has on users' trust, whereas Xie et al. (2019) discussed how medical professionals interact with the patient data. Thus, the differences in the questions in the articles made it once again difficult to make comparable observations for this category.

3.3.1.4 What questions

In total, five studies asked what-related questions. Two of the studies focused on discussing the positive and negative effects explanations might have on end-users' eyes. However, their findings were quite general and did not address precisely the same issues. Chazette and Schneider (2020) asked what kind of current perceptions end-users have on the need for explanations and what kind of advantages and disadvantages there are for having explanations. Dodge et al. (2019) also studied the advantages and drawbacks explanations might have "in supporting fairness judgment of ML systems". Chazette and Schneider (2020) found out that users need explanations. There are both advantages and disadvantages concerning three fields: informativeness and understandability of the system, usability, and relationship with the system. When it comes to Dodge et al. (2019), their main conclusion regarding advantages was that some fairness issues, such as model-wide fairness problems, might be revealed more effectively by different types of explanations.

The question settings for the last three articles in this category were once again very heterogeneous. Two of the studies discussed the role of collaboration and interaction but from different perspectives. Whereas Holstein et al. (2019) explored in their study what are the roles of human versus AI interaction in students' and teachers' opinion, discussed Oh et al. (2019) what factors are associated with the process in which users and AI collaborate in creative tasks. Ngo et al. (2020) discussed mental models: the mental models' end-users have on recommender systems and the extent to which the systems are seen as transparent or controllable. Therefore, no other comparable observations between the articles could be made.

3.3.1.5 Other research questions

One study's research question, Eiband et al. (2018), could not be accurately categorized. The authors did not explicitly state what their underlying research question guiding them in conducting the study is. Their contribution was to provide an overview of a stage-based, participatory process in which a transparent user interface would be created in the context of a fitness coach application.

3.3.2 Theoretical lenses used to understand XAI

Only four of the studies had some sort of explicitly mentioned theoretical lens through which they conducted their empirical research process. Two of the studies written by the same research group (Weitz et al. 2019; 2020) used the LIME framework proposed by Ribeiro et al. (2016). The articles based on theories used are listed in the following table.

Article	**Theories used**
Dodge et al. 2019	Four types of explanations created by Binns et al. (2018)
Weitz et al. 2019; 2020	LIME framework (see Ribeiro et al. 2016)
van der Waa et al. 2020	ICM framework (proposed by the research group themselves)

Table 6 Theoretical frameworks used by the articles

Dodge et al. (2019) use the four types of explanations formed by Binns et al. (2018) to set for their empirical research. The study of Binns et al. (2018) was also included in this review. Dodge et al. (2019) study the fairness perceptions of people using an algorithm

that makes predictions about reoffend rates of various individuals who have previously committed crimes. The four explanation types Binns et al. (2018) identify are input influence explanations (how much influence the input has on the outcome), sensitivity-based explanations, case-based explanations, and demographic-based explanations (based on aggregate statistics of the outcomes for people in the same demographic as the user). It is worth noting that Binns et al. (2018) study XAI from the perspective of justice perceptions, whereas the domain area in Dodge et al.'s (2019) research is criminal justice. The four explanation types might, therefore, not be universally valid to other domain areas.

The research group of Weitz et al. (2019; 2020) uses in both of their studies included in the sample the so-called LIME (Local Interpretable Model-Agnostic Explanations) framework created by Ribeiro et al. (2016). In both of their studies, they study the impact of virtual agents and speech recognition in XAI interaction design. The LIME Framework is used to generate the visual explanations in their experiments. The LIME framework is a model-agnostic approach or instead of an algorithm that can be used in explaining predictions of classifiers in an interpretable way "by learning an interpretable model locally around the prediction" (Ribeiro et al. 2016).

In the article of van der Waa et al. (2020), the research group themselves proposes a framework and then uses it as a lens to their own empirical investigations. Their study discusses how interpretable confidence measures (ICM) are interpreted by users of decision-support systems and what properties ICM's users need. The authors propose a framework for ICM's and then test it in two user experiments. They denote that ICM's should have at least four properties: they should be accurate, predictable, transparent, and explainable.

Besides the theoretical frameworks the articles used, the hypotheses articles proposed were investigated to determine what underlying literature, theories, and frameworks were behind those hypotheses. Of the 29 articles, in 11 articles, the researchers constructed one or more hypotheses for their empirical research. The articles and their hypothesis or hypotheses are presented in Table 7 below.

Article	**Hypotheses presented**
Broekens et al. 2010	"Our hypothesis is that different actions require different types of explanations, i.e., an interaction effect exists between type of explanation and action on the perceived quality of an explanation."
Chazette and Schneider 2020	"H10: There is no difference in the answers with respect to the advantages of explanations between digital natives and digital immigrants. H20: There is no difference in the answers with respect to the disadvantages of explanations between digital natives and digital immigrants."

Cramer et al. 2008	"H1: Users are more likely to accept a user-adaptive recommender system with a more transparent decision-making process. H2: Users will have more trust in a system with a more transparent decision-making process. H3: Users will perceive a user-adaptive recommender system with a more transparent decision-making process as more competent. H4: Understandable explanations of the reasons for a particular recommendation will increase acceptance and trust more than other types of transparency features."
Dodge et al. 2019	"Our hypothesis is that given local explanations focus on justifying a particular case, they should more effectively surface fairness discrepancies between cases."
Ehsan et al. 2019	"This study – seeks to confirm the hypothesis that humans prefer rationales generated by each of the configurations over randomly selected rationales across all dimensions."
Lim and Dey 2009	"We hypothesize that there are different types of information users are interested in, for different context-aware applications, and different situations." "We hypothesize that: (i) when asked specifically about whether they want an intelligibility type (heretofore called solicited information demand), users should reflect the same demands (elicited information demand) as that of experiment 1; and providing explanations for demanded intelligibility type will (ii) increase application satisfaction, and (iii) increase user rating of that intelligibility type."
Lim et al. 2009	"We hypothesize that different types of explanations would result in changes in users' user experience: understanding of the system and perceptions of trust and understanding of the system." "H1: Why explanations will improve user experience over having no explanations." "H2: Why Not explanations will (a) improve user experience over having no explanations, but (b) will not perform as well as Why explanations." "H3: How To or What If explanations will (a) improve user experience over having no explanations, but (b) will not perform as well as Why explanations."

Schrills and Franke 2020	"H1.1) Visual explanations of classifications based on the input stimulus lead to higher trust in the system. H1.2) Visual explanations of classifications based on the input stimulus lead to higher observability of the system. H2.1) Counterfactual explanations lead to higher trust than omni-explanations. H2.2) Counterfactual explanations lead to higher observability than omniexplanations. H2.3) Counterfactual explanations are rated as more understandable than omniexplanations."
Weitz et al. 2020	"For our hypothesis we assume a linear trend, which means that the general trust increases depending on the virtual agent group where the baseline group without agent has the lowest general trust score, followed by the text agent group, the voice agent group, and the virtual agent group with the highest scores in general trust."
Xie et al. 2019	"We hypothesize that human doctors will find a system more explainable when the system 'speaks the language' of a doctor and 'thinks like' a doctor."
Yin et al. 2019	"• [H1] The stated accuracy of a model has a significant effect on people's trust in the model before seeing the feedback screen. • [H2] The stated accuracy of a model has a significant effect on people's trust in the model after seeing the feedback screen. • [H3] The amount at stake has a significant effect on people's trust in a model before seeing the feedback screen. • [H4] The amount at stake has a significant effect on people's trust in a model after seeing the feedback screen. • [H5] The stated accuracy of a model has a significant effect on people's trust in the model before seeing the feedback screen, regardless of its observed accuracy. • [H6] The stated accuracy of a model has a significant effect on people's trust in the model after seeing the feedback screen, regardless of its observed accuracy. • [H7] After seeing the feedback screen, the observed accuracy of a model has a significant effect on people's trust in the model, regardless of its stated accuracy."

Table 7 Hypotheses presented by the articles

The application areas and hypotheses of the articles were as well various. Two groupings based on the hypotheses could be made. In the first one, three studies looked into the need for different explanations in different situations, whereas the other group studied visualizations in explanations. After discussing these groups, the theories referred to in the other studies are also discussed.

Broekens et al. (2010) base their hypothesis about different actions requiring different kinds of explanations on a previous study of Harbers et al. (2009) discussing the use of virtual training systems. In their study, it was found in different situations, distinct explanations were preferred. Broekens et al. (2010) also mention as their motive that Harbers et al. (2009) showed that in complex, action-requiring situations, humans could provide explanations that include only a couple of mental concepts.

Lim and Dey provide almost the same hypothesis in their 2009 study. They hypothesize that users are interested in different information in different situations and in using different context-aware applications. When forming their hypothesis, they refer to four previous studies. While they acknowledge that previous studies (Gregor and Benbasat 1999; McGuinness et al. 2007; Glass et al. 2008) have identified factors that users would like to have when using "adaptive agents" and knowledge-based systems, they see, referring to Abowd et al.'s (2002) study, that those systems differ from context-aware applications that are meant for more comprehensive, everyday use in ordinary life. Thus, want to explore what information needs would arise in these situations. In another study, from the same authors, Lim et al. (2009), they make a slightly different hypothesis. They presume that different types of explanations would change the user experience of the users', affecting their understanding and trust towards the systems. They refer again to Gregor and Benbasat's (1999) study and then to five types of intelligibility questions adapted from Dourish et al. (1996). While some studies (Ko and Myers 2004; Myers et al. 2006) have provided explanations to these intelligibility questions, the researchers see no evidence about their effectiveness, thus motivating their hypotheses.

When it comes to visualizations, Schrills and Franke (2020) propose four hypotheses in their study that explore the impacts of visual explanations on the AI systems' trustworthiness and observability. They motivate their hypotheses by referring to previous studies of Kulesza et al. (2012) and Bigras et al. (2018), in which there were found positive effects when visual explanations were presented. But to their knowledge, there are no studies that would compare the effects of "structurally different prototypical visual explanation approaches", thus, the setting for their study.

Continuing from the visualizations, as their hypothesis, Weitz et al. (2020) assume a linear trend, according to which the users' trust towards the AI system increases when adding explanation types, starting from plain text and concluding to the virtual agent. They refer to two previous studies when forming their hypothesis, De Graaf and Malle's (2017) and Van Mulken et al.'s (1998) studies. The previous study suggested that humans

approach explanations as if another human presented them, and that is why explanations should be conceptually and linguistically more human-like. In contrast, the latter study found out that personified agents improve the users' capabilities of processing technical information. They also refer to their previous study (Weitz et al. 2019), in which the use of virtual agents was already explored.

The other studies and their hypotheses were heterogenous, so they could not be grouped in a meaningful way. However, they all discussed different aspects that affect users' trust.

Chazette and Schneider (2020) hypothesize about digital natives and digital immigrants and their perceptions about the advantages and disadvantages of explanations. They mention two studies as motivation in defining their own hypotheses: Prensky's (2001) study and Hoffmann et al.'s (2014) study. Contrary to two of the prior studies they refer to, which see that digital natives relate to technologies differently and trust them more quickly than older generations, they hypothesize that there are no differences in digital natives' and digital immigrants' answers.

Cramer et al. (2008) then study the impact of transparency on art recommender systems. They provide a variety of literature as an inspiration to their hypotheses. They hypothesize that users would see transparent recommender systems as more acceptable, trustworthy, and competent and that understandable explanations will enhance acceptance and trust more than "other types of transparency features". Because there are many articles to be referenced, and, on average, they are more than fifteen years old, those studies are not presented in detail in this study.

In their hypothesis, Dodge et al. (2019) propose that local explanations should surface fairness discrepancies between cases more easily than other explanations since those focus on particular cases. They discuss the *case-specific disparate impact* and refer to the prior studies of Calders and Žliobaitė (2013) and Grgic-Hlaca et al. (2018) that discuss the discriminative decision procedures and fairness in algorithmic decision making. The disparate impact means that two individuals having the same kind of statistics and records except their race would get a different kind of treatment from judicial algorithms.

Xie et al. (2019) study the medical professionals' reasoning about clinical decision support systems and hypothesize that doctors would prefer explanations "speaking their [professional] language". They state that as a limitation of the current literature, there is no knowledge about XAI approaches supporting medical professionals' understanding and refer to the previous study of Krause et al. (2016).

Lastly, Yin et al. (2019) study the effects of the model's stated accuracy on user trust. They refer to a wide variety of literature in their prior work section, from which the studies of Kennedy et al. (2018) and Yu et al. (2016) can be raised as the most important motivators for hypotheses, as they are the only ones dealing with models' stated accuracy.

For one article, the hypotheses could not be directly linked to any of the previous literature. This case was the article of Ehsan et al. (2019). The hypothesis's formation was made in the section discussing the study's research method, and no reference was made to previous literature. However, the article does present prior literature, but it is impossible to draw direct similarities to the hypothesis from them.

3.3.3 Synthesis of the research question and theoretical approach analysis

Some concluding remarks can be made based on the research question analysis and the theoretical approach analysis. Based on which-questions, it was found out that users want and need certainty information (how certain the prediction of the AI system is), and that findings based on whether-questions findings would seem to imply that the effects of current explainability solutions to users' trust towards AI systems are pretty small or even non-existing.

When it comes to how-questions, it was seen that case-based explanations were found to be less fair than, for example, global and sensitivity-based explanations. Visualizations and especially virtual agents would seem to have some positive effects on users' trust. In the last category, what-questions, any comparable observations could not be made.

When it comes to theoretical approaches, it was surprisingly found out that only three studies took some kind of direct theoretical lens when conducting their empirical research. However, this might be due to the research field's relative freshness and that it is not a common practice in computer science. In this context, the most interesting framework was the LIME framework, which was often mentioned among the other sample articles.

Examination of the hypotheses also revealed more interesting, previous literature. However, several references to the same articles were not found behind the hypotheses' theoretical background due to the articles' diversity. However, the articles and their hypotheses provide interesting starting points for further research, such as visualization of explanations, user trust, and explanations' situational suitability.

4 DISCUSSION

In this section, the results section's findings are combined with the scientific background presented in the introduction. The key findings are stated, and the significance of the results and how the results respond to the research problems posed are considered.

Also, the limitations of the study and of the articles studied are presented, and their potential implications for the conclusions of the study are discussed. Finally, potential areas for further research are discussed based on the thesis and the sample articles' own observations.

4.1 Key findings

The descriptive key findings show that this study's sample was, due to the relative novelty of the research field, relatively new as expected. The majority of the articles had been published in 2018 or later. A large part of the articles combined qualitative and quantitative methods, while interviews, surveys, and user experiments were the most popular research methods among the sample.

Regarding stakeholder groups, it was found that in the eyes of the developers, users need explainability to make more informed decisions and to get a grasp of the capabilities of the system. All three articles (Holstein et al. 2019; Hong et al. 2019; Liao et al. 2020) discussing developers' needs also hint that the practitioners would need better tools to support them in the quest of providing explainability.

The key findings regarding domain experts included that the medical professionals tended to prefer receiving rather too much information than too little (see Cai et al. 2019; Wang et al. 2019; Xie et al. 2019), which was in stark contrast with fraud detection specialists' preference to have just the right amount of selective information (Cirqueira et al. 2020). Medical professionals prefer meticulousness in their decision-making, while speed is the primary quality criterion in decision-making for fraud detection professionals. This finding highlights how the needs for explainability vary between different domains, depending on the unique needs of each domain expert group.

In the last of the stakeholder groups, end-users, the significant themes included factors that increased users' trust and understanding towards AI systems. Due to the heterogeneity of the articles, there were only a few generalizable findings. Compared to developers and domain expert groups, providing synthesis to this cluster was more complicated because the articles' application areas varied wildly. As some of the most precise results, the case-based explanations were seen generally as less fair than other explanation styles, whereas sensitivity-based styles were the most effective ones (see Binns et al. 2018; Dodge et al. 2019). The end-users' prior conceptions about AI were as well described as

highly determinative to their judgment of fairness (see Binns et al. 2018; Dodge et al. 2019). Explanation interfaces also seemed to increase users' understanding of the model but not it trustworthiness (Cheng et al. 2019). Another common finding was that explanations must be formed so that the language and the context are relatable to the users and that they do not overwhelm users (see, e.g., Eslami et al. 2018; Ehsan et al. 2019).

The third and four analysis lenses took in this thesis were the research question-based and theoretical approach-based analyses. When it comes to the research question analysis, it was found out that users want and need certainty information (see Lim and Dey 2009; Chazette and Schneider 2020) and that findings based on whether-questions findings would seem to imply that the effects of current explainability solutions to users' trust towards AI systems are pretty small or even non-existing (see Cramer et al. 2008; Cheng et al. 2019; Yin et al. 2019). Through this analysis, case-based explanations were again found out to be the most unfair type of explanations (see Binns et al. 2018; Dodge et al. 2019) and that visualizations and especially virtual agents would seem to have some positive effects on users' trust (see Schrills and Franke 2020; Weitz et al. 2020).

When it comes to theoretical approaches, only three theoretical frameworks were found, through which the articles mirrored their empirical research. The most promising one of these frameworks concerning practical implications was the LIME framework (see Ribeiro et al. 2016). The examination of the hypotheses also revealed some more exciting literature. However, due to the articles' diversity, no previous theories to which two or more articles among the sample would have referred to were not found.

4.2 Answering the research questions

To repeat, the research question for this thesis, based on the research area and research gap, was identified as: *how do users' perceptions of transparency, explainability, and trustworthiness of AI manifest in the HCI literature?*
In support of the leading research question, the additional research questions used were:

- To whom is the XAI for? What are its stakeholders?
- What factors make AI transparent, explainable, or trustworthy for these target audiences?
- Do different audiences differ in their perceptions of the explainability of AI?
- What kind of different needs target audience groups have for explainability?

There are clear answers to the first additional research question: to whom explainability is for, whom to explain? In this thesis, three groups have been identified, to whom the current research tries 1) either to provide explanations (end-users, domain experts) or 2) seeks to ask their opinions and practices about explainability (developers).

The end-user group was the most significant group with 20 studies. The developer-group focuses on the perspectives of the practitioners and their teams creating and developing AI solutions. The domain experts group consisted of studies focusing on specialists in their field (another than AI), such as medical professionals, and the third group, end-users, consists of studies focusing on consumers, ordinary people, and other groups that use artificial intelligence solutions but are not actively involved in their design processes.

Extant literature's heavy focus on end-users implicates a lot of curiosity about how users perceive and understand XAI. The heterogeneity of the sample articles reflects this interest well, as the articles in this sample included, for example, perspectives from the fields of recommender systems to autonomous systems, from gaming to dating applications, and from tutoring systems to criminal justice. Another perspective that received significant attention among the sample articles was the perspective of medical professionals. This is natural, as AI systems are already used to some extent in the medical field and have great potential as a tool for better treatment of diseases.

To the second additional research question (what factors make AI transparent, explainable, or trustworthy in these target audiences' eyes), some findings emerged. Some domain expert groups, such as medical professionals, need extensive transparency, while other groups (in this case, fraud detection specialists) prefer to have explanations that support fast decision making. In light of the end-user group's findings, some explanations were found to be more transparent than others. The end-users' prior conceptions towards AI were found to define their perceived trust towards AI systems in general. In conclusion to this additional research question, there are some takeaways, but very far-reaching conclusions cannot be drawn due to the sample's heterogeneity. This is also true concerning the third additional research question (the differences between target audience groups' perceptions of explainability). Of course, the developer group differed from the other two groups as developers influence explainability rather than are its targets. However, differences between domain experts and end-users are much harder to find. The domain users' viewpoint to explainability depends much on the domain, and there are significant differences within the end-user cluster as well based on the application area. The most apparent difference is that domain experts see explainability as a tool that can help them achieve their goals, while end-users seem to perceive it as a feature.

To the fourth additional question (what kind of different needs target audience groups have for explainability?), some remarks can be made. The developers would need more tools supporting them in creating explainability; the domain experts, depending on the domain, want to have access to the right amount of background data and the model's inner workings. In contrast, end-users want to have explanations that "speak their language", are not creepy, nor are too detailed. The users, in general, want to have information "at the right time", and explanations should support their existing mental models. Especially

the end-users do not want to have case-based explanations since they consider it unfair to make any predictions or decision based on some particular, historical case.

Lastly, regarding the main research question (how users perceive the transparency, explainability, and reliability of artificial intelligence), some conclusions can be drawn. Users seem to appreciate explainability and want to have explanations, whether they are end-users or domain experts. Developers even see providing explanations as to their duty. Many of the current AI systems are not seen as transparent, and users often provide contrasting opinions about how the experience should be improved. When it comes to understandability, the explanations seem to affect users positively, but evoking trust and transparency seem to be much more complex tasks. Regarding reliability, users' prior conceptions about AI seem to affect their judgment, so that despite their excellent features, it is difficult for AI systems to gain their trust. In the light of these findings, users still feel somewhat skeptical about XAI's potential.

4.3 Implications of findings

The findings discussed above have some broader implications for the scientific debate and to the surrounding society. In this section, some of those implications are presented and discussed.

4.3.1 Implications for research

This study's most significant scientific contribution is the synthesis of the empirical XAI studies that have already been published. This thesis provides an overall picture of where, in terms of user experience, the research field of XAI is now and thus helps researchers target their future research better. XAI research is still new and fragmented, and this thesis provides clarity to the situation. The work brings together recent research and shows that, while studies are discussing the HCI perspective, the studies usually focus on some particular technical solutions and their effects on users' perceived trust and understanding. The studies are pretty point-like and experimental, and broader, longer-term empirical studies from the HCI-perspective would seem to be still lacking from the research field. The findings, however, open avenues for new research and provide some underpinnings of the theories currently used in XAI research. The findings also provide a window into XAI developers' world and needs that have not been over-explored in the previous literature.

While the findings of this study show that while the literature is still fragmented, the research field has, however, started to align under the concept of XAI. However, for

example, Brennen (2020) noted that the variation of terms is still considerable in the terminology used by experts and stakeholders in the field and even in scientific literature, as noted in this thesis's course. To promote shared understanding and ensure the discoverability of the studies, it would be desirable for researchers to use at least the keywords "machine learning" and "explainability" in addition to XAI in their future research.

Regarding the publication venues, the most popular one among the sample was the Proceedings of the CHI Conference on Human Factors in Computing Systems, which is reasonable since it is one of the biggest and most prestigious in the research field. The articles' publication was also concentrated on the conference proceedings among the sample, which is typical for computer science as a scientific discipline. Based on their data collection, the articles' geographical distribution was heavily concentrated on the United States, which cannot be considered particularly surprising, as the United States is the leading country in the research field. These findings imply that the Anglo-American science sphere heavily dominates the research field. More extensive XAI research from the rest of the world would be needed in the future to complement its perspectives.

Another scientific implication is that this thesis supports the findings of the literature review of Arrieta et al. (2020) by explicitly bringing the HCI perspective into the discussion. In addition to literature reviews that outline the XAI field and its trends more generally, there has been a need to examine the issue specifically from an HCI perspective – a challenge that this thesis seeks to address on its part.

One theoretical implication discussed as a deficiency in the current literature (see, e.g., Holzinger et al. 2020) but noted even based on this study's observations was the lack of quality criteria and quality in general in explanations. For example, it is noted that while there are several methods to produce explainability, the research community does not have any shared formal standards or evaluation methodologies for explanation quality (Binns et al. 2018). This lack is logical since explanations depend on the context since there are different users, different needs, and functions in different situations. However, it would be interesting to see what kind of general quality criterion there could even exist. Could some thumb rules could be used no matter what the situation is? Exploring this topic could potentially provide interesting tools for future developers.

Another implication of scientific interest is the algorithm auditing techniques or, to be more exact, the lack of them, which is also one of the AIGA research project's critical topics. The lack of algorithm auditing methods was noted in the studies of Holstein et al. (2019) and Eslami et al. (2018). Based on these findings, domain-specific and system-level auditing methods would be needed, both of which would be fields for new research. More broadly, more tools helping practitioners create better algorithms would be needed, which would be a fruitful area for new research.

Continuing with the developer theme, the findings imply that there would be a need for research looking detailly into what factors prevent practitioners from achieving their

desired level of explainability when creating ML models. These themes were partly discussed by the articles focusing on developers' viewpoints (Holstein et al. 2019; Hong et al. 2020), but both articles focused more on the developers' needs. Of course, the topics are closely intertwined, challenges cannot be solved without answering the developers' needs. However, the topic would offer exciting opportunities from the perspectives of organizational research, competency research, and social relations research.

The findings have some implications that hint that XAI should be treated as a cooperative process. In the end-user section, the study of Oh et al. (2018) illustrated examples of how in human-AI art co-creation, the users wanted to control the art creation and pushed the practical work to the shoulders of the AI. In the developer group, there are some implications in the findings of Holstein et al. (2019) and Hong et al. (2020), that the developers have a comprehensive need for collaboration and would like to have even between-organization collaboration and share practices for creating better AI. Wang et al. (2019) have also proposed (referring to Miller 2017) that explanations are distinctively social as their nature, and that explanations should be seen as collaborative conversations. One exciting research area for the future could be thus discussing XAI as a collaborative process. These collaborative and conversational aspects would provide fruitful and multifaceted research avenues.

Expanding to end-users' perceived needs, it was found in Chazette and Schneider's study (2020) that explanations should be presented only when the user demands them. Further research would be needed to support this finding, as the article was the only one in which such a strong position was presented for showing explanations. Continuing this finding towards defining in more general situations where users would like to have explanations and forming theories based on them would provide interesting knowledge to many fields.

One last theoretical implication is that the users' mental models in the context of XAI have not been explored very widely. For example, Ngo et al. (2020) propose that future research should explore the transfer of mental models within different platforms. In contrast, Schrills and Franke (2020) propose that future research should more closely examine the development of mental models in the XAI context. Continuing from the same theme, one potential future research area would be looking into users' prior conceptions about AI and their trust-building processes. Previous studies, such as Binns et al. (2020) and Dodge et al. (2019), have highlighted the effect users' prior conceptions have on AI are considerable and can hinder the trust-building process between them and the AI. Looking more in detail into this phenomenon could bring important considerations to the design of more user-friendly and trust-inspiring XAI systems.

4.3.2 Implications for practice

Since many hopes have been placed on AI and its development, it is clear that the results of this thesis are of interest to a wide range of AI stakeholders, spanning from banks and insurance companies to online stores, legislators, and other authorities. Many AI systems and applications, according to studies, are coming into use but are not yet available to the public. How should this change be reflected at the level of AI systems design and the level of legislation? The following paragraphs introduce some implications for these themes.

The companies and developers who create new artificial intelligence solutions can benefit from the findings of this thesis by gaining a broader overview of what has recently happened in the field of XAI research from the HCI perspective. From a practical point of view, the findings provide an understanding of what users view as explainable, thus facilitating the user-friendly design of future artificial intelligence solutions. Based on the findings, companies can strive to develop their processes and ways of collaborating and the necessary tools to improve explainability. One concrete example of this involves the presentation styles of explanations. Based on several articles' findings (see, e.g., Binns et al. 2018; Dodge et al. 2019), case-based explanations were seen by users as less transparent and less reliable than, for example, sensitivity-based explanations. For example, developers of autonomous systems, such as cars, should take this fact into account when developing explanations for their systems. Users prefer to receive guidance based on more general statistics and averages than based on a single previous case, even if that case would reflect the system's recommendations in general. As a design principle, one could suggest that case-based explanations should be mainly used to support other explanations and situations where other explanation styles cannot be used.

Another implication identified by the study relates to recommender algorithms and their power. Based on the findings of a couple of articles (see, e.g., Cramer et al. 2008; Ngo et al. 2020), users perceive recommender systems often as opaque and do not understand how they work. Because of the business logic, it is understandable that the operation of business secrets, such as recommendation algorithms, is not unduly open, but this is also directly reflected in the weakening of the user experience and user confidence. Developers of AI-based recommendation systems should seek to address this problem in the future to strengthen the user experience, the usefulness, and the users' trust towards the systems. The same problem applies to the field of online advertisement algorithms, which, based on the findings of Eslami et al. (2018), whose way of selecting and targeting advertising to each user is not understood by ordinary end-users and may even arouse feelings of anxiety in them.

There is one practical implication that stemmed from the papers exploring the developer stakeholder group. The developers brought out in a few papers that there is a need

for better tools supporting the XAI development and especially for collaboration between the practitioners and companies within the field so that best practices could be shared more openly (see, e.g., Holstein et al. 2019; Hong et al. 2019; Liao et al. 2020). This cooperation would benefit the whole society, as with the help of shared information, XAI solutions would become increasingly better. Therefore, companies in the sector could form clusters and collaborative projects in which good practices and explanations about XAI would be shared.

From the point of view of society, the thesis contributes to providing some preliminary answers on how AI could be made more transparent and thus more sustainable and usable in the eyes of a larger audience. The findings can be channeled, for example, to the design of public administration AI systems. To name a few, the studies of Binns et al. (2018) and Dodge et al. (2019) discuss the perceived justice and transparency of AI systems in the field of criminal justice, whereas Cheng et al. (2019) study the effects of algorithm use in university admissions. Their findings show that regarding users' trust, their prior conceptions about the suitability of AI in the domain may undo all the benefits of the system and improvements made to it if users consider the task type fundamentally unfit for algorithmic decision making. This finding must be considered when designing future AI solutions in the public administration field. For now, AI would not seem appropriate for all public administration tasks, at least if the aim is to gain users' trust.

This study's findings also have implications for the medical and nursing sector, its practitioners, public administration and private actors, and the technology companies operating in the field. From the findings, it would seem clear that medical professionals want to have comprehensive access to the data the AI system uses and information about its underlying assumptions, such as the theoretical basis on which AI has been trained (see, e.g., Cai et al. 2019; Xie et al. 2019). Without the opportunity for medical professionals to gain a deeper understanding of AI's decision-making process, its potential in the treatment of diseases is likely to remain untapped. This factor should be considered by companies and researchers developing medical systems in their future work.

Regarding the role of AI legislation, the articles in this study provide little further answers. The advent of GDPR legislation has been referred to in several papers (see, e.g., Chazette and Schneider 2020). Still, the broader implications for future legislation and the users' right to receive explanations remain unaddressed. This finding is one premise that future legislators, on the one hand, and AI researchers and jurists, on the other, should consider together.

4.4 Limitations

4.4.1 Limitations of the research process

There are some potential limitations to this study, which will be discussed in this section. The possibility of biases during the data collection process cannot be excluded. One of the main limitations of this study is that the inclusion and exclusion of the articles to the final sample was done by only one person, the writer of this thesis. The article search was thus not done in duplicate. A common practice in systematic literature reviews that the screening of the articles is done by at least two scholars so that their decisions can be compared and borderline cases discussed. Often a reliability score is provided to visualize the similarity between the work of the researchers. This practice was not possible in the context of this study since this is a master's thesis designed to demonstrate the skills acquired by its author during his studies.

The heterogeneity and relevant novelty of the research area may have induced some limitations to this study. First, since the scientists in this field have used a variety of synonyms describing XAI and explainability, it may well be that during the research process, some terms have been understood differently than the original writers have intended by the author of this thesis.

Another limitation stemmed from the choice of the search terms of this study. As the research field is still very new and heterogeneous in its use of terms, there might have been some relevant studies that did not come up in the database searches simply because the researcher or research group used utterly different terms than the author of this thesis. One significant lack in this study's search strings was identified as the process proceeded: "machine learning" was not included in them. However, the weight of missing a keyword was diminished by the fact that snowball sampling was conducted to find additional articles. Indeed, through snowball sampling, several articles were found that discussed the research field mainly by talking about machine learning, and XAI was not necessarily mentioned at all.

The longitudinal effects may cause some limitations to this study. Since the research period was relatively short due to the common interests of the research group and the thesis writer advancing rapidly, it can affect this thesis's results. As the writer of this theses does not have prior academic or practitioner knowledge about AI or writing systematic literature reviews, it might affect the research process and the results of this study. Another limitation to the thesis is related to the articles' exclusion criteria: since the article search was concluded on October 20, 2020, no further articles published after that were included in the review. This temporal limitation is essential to note since new articles about XAI are emerging in the databases presently almost daily.

Another limitation of the research process was that the articles were not analyzed based on their application area due to the research process's tight schedule. This analysis could have been done by grouping articles based on their research questions into groups investigating similar issues, for example, instead of or in addition to a question-based approach. This approach can still be utilized in further research based on this thesis.

4.4.2 Limitations within and across studies

There are some limitations and biases within the reviewed articles that are worth mentioning. The geographic extent of studies among the sample is one limitation of this study. Based on the university affiliations of the writers and countries from where the empirical data was collected, the sample was heavily concentrated on the Anglophone (United States, United Kingdom, Ireland, and Canada) and Germanic (Germany, the Netherlands, Austria) countries. The rest of the world had only a few studies among the sample. This lack might affect the reliability of the results in countries and contexts in which the cultural and social conditions are different from those in these countries, dominating the sample of this study.

One possible limitation to this study may be the biases within the empirical samples of the reviewed articles. Many articles among the review sample (such as Cheng et al. 2019; Dodge et al. 2019; Yin et al. 2019, etc.) used crowdsourcing to collect answers to their end-user surveys. The most used tool for this was Amazon’s crowdsourcing service, Mechanical Turk. Crowdsourcing might affect the reliability of the studies' results, as in these cases, the results reflect the opinions of the people who specifically use that service and cannot be generalized to a broader population.

One of the thesis's limitations is that the synthesis was partly hampered by the disciplines' fragmentation and different starting points. The application areas, target audiences, and even terms used in the articles varied very much; it is easy to say that the article sample was very heterogeneous. Therefore, it was not possible to make very extensive generalizations based on the observations, as it was often possible to present only findings of a couple of articles to support the observations. This limitation is a good reflection of how research into the XAI field from the HCI perspective is still in its infancy.

4.5 Future work

Several interesting future research topics could be explored based on the findings of this study. These observations in detail are discussed more broadly in section 4.3.1., which

discusses this study's theoretical implications. In this section, the areas future work should focus on are presented more broadly.

As noted before in section 4.3.1., future work should specifically look into topics, such as XAI as a collaborative process, explanation quality methodologies, algorithm auditing methods, developers' perceived XAI challenges, users' mental models, and their prior conceptions about AI. Thus, future research should examine the shaping of user experience to influence user attitudes toward AI and minimize the impact of past (negative) perceptions. In the light of the findings, research focusing specifically on developers' own perspectives and work is also needed so that more user-centric XAI solutions, in general, can be developed.

A lot of important work on AI development and XAI is done outside the research world; future work should explore the written and audiovisual material outside scientific publications. Important work on AI is constantly published in blog posts, corporate websites, and YouTube videos. They do not directly convey to the academic world interested in the material published in scientific publications. Exploring these unscientific platforms would complement the existing literature with perspectives that would otherwise remain hidden.

One area of research that has, based on this study, so far remained almost entirely unaddressed in terms of empirical research relates to the banking and insurance industries. Although artificial intelligence solutions are already quite common in the banking and insurance sector as one of the first industries (see, e.g., Digalaki 2021), only one of the articles in the sample of this study discussed the financial sector (Cirqueira et al. 2020), even then from the point of view of fraud detection professionals. Future research could focus on using XAI in the banking and insurance industry from the perspective of their customers and customer experience.

Another surprising shortcoming in the context of this study was the lack of research on autonomous systems from the HCI perspective. Only two studies (Broekens et al. 2010; Chazette and Schneider 2020) had empirically examined users 'experiences of those systems' explanations. When self-driving cars and other autonomous AI systems are often referred to as solutions for the future, this theme could also be studied more from the HCI perspective.

Future research can also test the reliability of the findings of this study. Future researchers could duplicate this research or parts of it after some time has passed to see what kind of differences newer articles would provide. As XAI research is constantly evolving and new articles are published on an almost daily basis, it is not far-fetched that the results of this study might look quite different in a year. As the need for XAI research continues to grow, more and more comprehensive research on it will also emerge in the future.

5 CONCLUSIONS

The purpose of this study has been to outline through an explorative, systematic literature review the current discussions and themes found in the Explainable AI (XAI) research literature from the human-computer interaction (HCI) perspective. In total, 29 articles that concluded an empirical study into XAI through the HCI perspective were discovered. The research was carried out by using the PRISMA method for systematic literature reviews. The articles were analyzed through four lenses: their descriptive statistics, their stakeholder groups, and significant findings within the audience groups, and through the research questions and theoretical approaches, the articles took. This study aimed to determine what factors made users consider XAI trustworthy, explainable, or reliable and to whom the XAI research was intended for.

This thesis's most immense contribution is to provide a synthesis of the extant empirical XAI literature from the HCI perspective, which previous studies have rarely brought together. It was found that domain experts' needs towards XAI vary significantly between domains, whereas developers would need better tools to support them in creating XAI systems. The end-users, on their part, considered case-based explanations unfair and wanted to have explanations that "speak their language". The research question analysis concluded that the effect of current XAI solutions on users' trust towards AI systems is relatively small or even non-existing. Case-based explanations were again found out to be the most unfair type of explanations. Visualizations, however, had some positive impacts on users' perceived trust. The findings bring some much-needed understanding of what users view as explainable. Based on the findings, companies and developers can develop their processes, and researchers can provide new theories to help facilitate the ever-increasing collaboration of AI systems and humans. This study also sheds light on the theoretical background the articles in the field of XAI have. As the XAI field is practice-oriented, the studies' theoretical contributions and the number of theoretical lenses used were both found out to be relatively low. The lack of theoretical frameworks and the lack of tools desired by developers need tells how critical and needed XAI research and its applications are at the moment.

As a future outlook, researchers, developers, and stakeholders in the XAI community can use the findings to further their initiatives towards more explainable AI systems. Based on this thesis, future research avenues, such as XAI as a collaborative process, explanation quality methodologies, algorithm auditing methods, developer challenges, users' mental models, and prior conceptions about AI, could be further investigated. The thesis also provides a good starting point for other researchers and practitioners to overview the current user-centric XAI literature. As the need for and the publication of XAI-related literature is increasing at an ever-accelerating pace, it will be a fruitful research area for many scientists to come.

REFERENCES

Abowd, G. D. – Mynatt, E. D. – Rodden, T. (2002) The Human Experience [of Ubiquitous Computing]. *IEEE Pervasive Computing*, Vol. 1(1), 48–57.

Adadi, A. – Berrada, M. (2018) Peeking Inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, Vol. 6, 52138–52160.

Baker, S. E. – Edwards, R. (2012) How Many Qualitative Interviews is Enough? Expert Voices and Early Career Reflections on Sampling and Cases in Qualitative Research. *National Centre for Research Methods Review Paper*, 1–42.

Barredo Arrieta, A. – Díaz-Rodríquez, N. – Del Ser, J. – Bennetot, A. – Tabik, S. – Barbado, A. – Garcia, S. – Gil-Lopez, S. – Molina, D. – Benjamins, R. – Chatila, R. – Herrera, F. (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, Vol. 58, 82–115.

Bigras, E. – Jutras, M. A. – Sénécal, S. – Léger, P. M. – Black, C. – Robitaille, N. – Grande, K. – Hudon, C. (2018). In AI We Trust: Characteristics Influencing Assortment Planners' Perceptions of AI Based Recommendation Agents. In *Proceedings of the 2018 International Conference on HCI in Business, Government, and Organizations*, eds. Nah, F.H. – Xiao, B., *Lecture Notes in Computer Science*, Vol. 10923, 3–16. Springer, Cham.

Binns, R. – Van Kleek, M. – Veale, M. – Lyngs, U. – Zhao, J. – Shadbolt, N. (2018) 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14, Montréal, Canada.

Bossmann, J. (2016) Top 9 Ethical Issues in Artificial Intelligence. *World Economic Forum*. <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/>, retrieved 8.10.2020.

Brennen, A. (2020) What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7, Honolulu, United States, 2020.

Broekens, J. – Harbers, M. – Hindriks, K. – Bosch, K. – Jonker, C. – Meyer, J-J. (2010) Do You Get It? User-Evaluated Explainable BDI Agents. In *Proceedings of the*

8th German Conference on Multiagent System Technologies, 28–39. Springer, Berlin/Heidelberg.

Bunge, M. (1967) Scientific Research: Strategy and Philosophy. Volume 1: The Search for System. Springer-Verlag, Berlin/Heidelberg.

Bussone, A. – Stumpf, S. – O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *Paper presented at the 2015 International Conference on Healthcare Informatics*, Dallas, United States, October 21–23, 2015, 160–169.

Cai, C. J. – Winter, S. – Steiner, D. – Wilcox, L. – Terry, M. (2019) "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction*, Vol. 3, No. CSCW, 1–24.

Calders, T. – Žliobaitė, I. (2013) Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In: *Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics*, eds. Custers B. – Calders T. – Schermer B. – Zarsky T., 43–57. Springer, Berlin, Heidelberg.

Chazette, L. – Schneider, K. (2020) Explainability as a Non-Functional Requirement: Challenges and Recommendations. *Requirements Engineering*, Vol. 25(4), 493–514.

Cheng, H. F. – Wang, R. – Zhang, Z. – O'Connell, F. – Gray, T. – Harper, F. M. – Zhu, H. (2019). Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12, Glasgow, United Kingdom, 2019.

Cirqueira, D. – Nedbal, D. – Helfert, M. – Bezbradica, M. (2020) Scenario-Based Requirements Elicitation for User-Centric Explainable AI. In: *Machine Learning and Knowledge Extraction - CD-MAKE* 2020, eds. Holzinger, A. – Kieseberg, P. – Tjoa, A. – Weippl, E., *Lecture Notes in Computer Science*, vol 12279, 321–341. Springer, Cham.

Cramer, H. – Evers, V. – Ramlal, S. – Van Someren, M. – Rutledge, L. – Stash, N. – Aroyo, L. – Wielinga, B. (2008) The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender. *User Modeling and User-Adapted Interaction*, Vol. 18(5), 455–496.

Deeks, A. (2019) The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review*, Vol. 119(7), 1829–1850.

De Graaf, M. M. – Malle, B. F. (2017). How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposium Series,* 19–26.

Digalaki, E. (2021) The Impact of Artificial Intelligence in the Banking Sector & How AI is Being Used in 2021. *Business Insider*. <https://www.businessinsider.com/ai-in-banking-report?r=US&IR=T>, retrieved 15.3.2021.

Dodge, J. – Liao, Q. V. – Zhang, Y. – Bellamy, R. K. – Dugan, C. (2019) Explaining Models: an Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285, Los Angeles, United States, 2019.

Doshi-Velez, F. – Kortz, M. – Budish, R. – Bavitz, C. – Gershman, S. – O'Brien, D. – Scott, K. – Schieber, S. – Waldo, J. – Weinberger, D. – Weller, A. – Wood, A. (2017) Accountability of AI Under the Law: The Role of Explanation. arXiv preprint, arXiv:1711.01134.

Došilović, F. K. – Brčić, M. – Hlupić, N. (2018) Explainable Artificial Intelligence: A Survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215, Opatija, Croatia, 2018.

Dourish, P. – Adler, A. – Smith, B.C. (1996) Organising User Interfaces Around Reflective Accounts. *Reflection'96*, 235–244.

Duan, Y. – Edwards, J.S. – Dwivedi, Y.K. (2019) Artificial Intelligence for Decision Making in the Era of Big Data–evolution, Challenges and Research Agenda. *International Journal of Information Management*., vol. 48(2019), 63–71.

Dwivedi,Y.K. – Hughes, L. – Ismagilova, E. – Aarts, G. – Coombs, C. – Crick, T. – Duan, Y. – Dwivedi, R. – Edwards, J. – Eirug, A. – Galanos,V. – Ilavarasan, P.V. – Janssen, M. – Jones, P. – Kar, A.K. – Kizgin, H. – Kronemann, B. – Lal, B. – Lucini, B. – Medaglia, R. – Le Meunier-FitzHugh, K. – Le Meunier-FitzHugh, L.C. – Misra, S. – Mogaji, E. – Sharma, S.K. – Singh, J.B. – Raghavan, V. – Raman, R. – Rana, N.P. – Samothrakis, S. – Spencer, J. – Tamilmani, K. Tubadji, A. – Walton, P. – Williams, M.D. (2021) Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. *International Journal of Information Management*, vol. 57(2021), 1–47.

Ehsan, U. – Tambwekar, P. – Chan, L. – Harrison, B. – Riedl, M. O. (2019, March) Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274, Los Angeles, United States, 2019.

Eiband, M. – Schneider, H. – Bilandzic, M. – Fazekas-Con, J. – Haug, M. – Hussmann, H. (2018). Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 211–223, Tokyo, Japan, 2018.

Eslami, M. – Krishna Kumaran, S. R. – Sandvig, C. – Karahalios, K. (2018). Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13, Montréal, Canada, 2018.

Ferreira, J. – Monteiro, M. (2020) What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Lecture Notes in Computer Science*, Vol. 12201, 56–73.

Ferreira, J. – Monteiro, M. (2021) The Human-AI Relationship in Decision-Making: AI Explanation to Support People on Justifying Their Decisions. arXiv preprint arXiv:2102.05460.

Finnish Advisory Board on Research Integrity (2012) Responsible Conduct of Research and Procedures for Handling Allegations of Misconduct in Finland. <https://www.tenk.fi/sites/tenk.fi/files/HTK_ohje_2012.pdf>, retrieved 2.3.2021.

Gilpin, L. – Bau, D. – Yuan, B. – Bajwa, A. – Specter, M. – Kagal, L. (2018) Explaining Explanations: An Overview of Interpretability of Machine Learning. In *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89, Turin, Italy, 2018.

Glass, A. – McGuinness, D. L. – Wolverton, M. (2008) Toward Establishing Trust in Adaptive Agents. *In Proceedings of the 13th International Conference on Intelligent User Interfaces*, 227–236, Gran Canaria, Spain, 2008.

Gomboc, D. – Solomon, S. – Core, M. – Lane, H. – Van Lent, M. (2005) Design Recommendations to Support Automated Explanation and Tutoring. In *Proceedings of the 14th Conference on Behavior Representation in Modeling and Simulation*, Universal City, United States, 2005.

Goodman, B. – Flaxman, S. (2017) European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, Vol. 38(3), 50–57.

Gregor, S. – Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, Vol. 23(4), 497–530.

Grgic-Hlaca, N. – Redmiles, E. M. – Gummadi, K. P. – Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference*, 903–912, Lyon, France, 2018.

Guidotti, R. – Monreale, A. – Giannotti, F. – Pedreschi, D. – Ruggieri, S. – Turini, F. (2019) Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems*, Vol. 34(6), 14–23.

Gunning, D. (2017) Explainable Artificial Intelligence (XAI). Technical report, *Defense Advanced Research Projects Agency* (DARPA), 2017.

Gunning, D. – Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, Vol. 40(2), 44–58.

Gunning, D. – Stefik, M. – Choi, J. – Miller, T. – Stumpf, S. –Yang, G. (2019). XAI - Explainable Artificial Intelligence. *Science Robotics*, Vol. 4(37).

Hacker, P. – Krestel, R. – Grundmann, S. – Naumann, F. (2020) Explainable AI Under Contract and Tort Law: Legal Incentives and Technical Challenges. *Artificial Intelligence and Law*, Vol. 28, 415–439.

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer*, Vol. 51(9), 28–36.

Harbers, M. – Van den Bosch, K. – Meyer, J. (2009) A Study into Preferred Explanations of Virtual Agent Behavior. In: *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, eds. Ruttkay, Z. – Kipp, M. – Nijholt, A. – Vilhjálmsson, H., *Lecture Notes in Computer Science*, vol. 5773, 132–145. Springer, Heidelberg.

Hennink, M. M. – Kaiser, B. N. – Marconi, V. C. (2017) Code Saturation Versus Meaning Saturation: How Many Interviews are Enough? *Qualitative Health Research*, Vol. 27(4), 591–608.

Hoffmann, C. P. – Lutz, C. – Meckel, M. (2014) Digital Natives or Digital Immigrants? The Impact of User Characteristics on Online Trust. *Journal of Management Information Systems*, Vol. 31(3), 138–171.

Hohman, F. – Head, A. – Caruana, R. – DeLine, R. – Drucker, S. M. (2019) Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13, Glasgow, United Kingdom, 2019.

Holstein, K. – Wortman Vaughan, J. – Daumé III, H. – Dudik, M. – Wallach, H. (2019) Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16, Glasgow, United Kingdom, 2019.

Holzinger, A. (2018) From Machine Learning to Explainable AI. In *Proceedings of the DISA 2018 - IEEE World Symposium on Digital Intelligence for Systems and Machines*, 55–66, Košice, Slovakia, 2018.

Holzinger, A. – Carrington, A. – Müller, H. (2020) Measuring the Quality of Explanations: the System Causability Scale (SCS). *KI-Künstliche Intelligenz*, Vol. 34, 1–6.

Hong, S. R. – Hullman, J. – Bertini, E. (2020) Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, No. CSCW1, 1–26.

Jha, S. – Sahai, T. – Raman, V. – Pinto, A. – Francis, M. (2019) Explaining AI Decisions Using Efficient Methods for Learning Sparse Boolean Formulae. *Journal of Automated Reasoning*, Vol. 63(4), 1055–1075.

Keneni, B. M. – Kaur, D. – Al Bataineh, A. – Devabhaktuni, V. K. – Javaid, A. Y. – Zaientz, J. D. – Marinier, R. P. (2019) Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles. *IEEE Access*, Vol. 7, 17001–17016.

Kennedy, R. – Waggoner, P. – Ward, M. (2018) Trust in Public Policy Algorithms. Working paper, 2018.

Kenny, E. M. – Ruelle, E. – Geoghegan, A. – Shalloo, L. – O'Leary, M. – O'Donovan, M. – Keane, M. T. (2019) Predicting Grass Growth for Sustainable Dairy Farming: A CBR System Using Bayesian Case-Exclusion and *Post-Hoc*, Personalized Explanation-by-Example (XAI). In: *Case-Based Reasoning Research and Development – ICCBR 2019*, eds. Back, K – Marling, C, *Lecture Notes in Computer Science*, vol 11680, 172–187. Springer, Cham.

Kirsch, A. (2018) Explain to Whom? Putting the User in the Center of Explainable AI. In *Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML*, Bari, Italy, 2017.

Kitchenham, B. (2004) Procedures for Performing Systematic Reviews. *Joint Technical Report, Computer Science Department, Keele University and National ICT Australia Ltd.*, 33(2004), 1–26.

Ko, A. J. – Myers, B. A. (2004). Designing the Whyline: a Debugging Interface for Asking Questions about Program Behavior. In *Proceedings of the 2004 SIGCHI Conference on Human Factors in Computing Systems*, 151–158, Vienna, Austria, 2004.

Krause, J. – Perer, A. – Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. *In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5686–5697, San Jose, United States.

Kulesza, T. – Stumpf, S. – Burnett, M. – Kwan, I. (2012). Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems*, 1–10, Austin, Texas, United States.

Kuwajima, H. – Tanaka, M. – Okutomi, M. (2019) Improving Transparency of Deep Neural Inference Process. *Progress in Artificial Intelligence*, Vol. 8(2), 273–285.

Li, Y. – Wang, H. – Dang, L. M. – Nguyen, T. N. – Han, D. – Lee, A. – Jang, I. – Moon, H. (2020) A Deep Learning-Based Hybrid Framework for Object Detection and Recognition in Autonomous Driving. *IEEE Access*, Vol. 8, 194228–194239.

Liao, Q. V. – Gruen, D. – Miller, S. (2020) Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15, Honolulu, United States, 2020.

Lim, B. – Dey, A. (2009) Assessing Demand for Intelligibility in Context-Aware Applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, 195–204, Orlando, United States, 2009.

Lim, B. – Dey, A. – Avrahami, D. (2009) Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the 2009*

SIGCHI Conference on Human Factors in Computing Systems, 2119–2128, Boston, United States, 2009.

Longo, L. – Goebel, R. – Lecue, F. – Kieseberg, P. – Holzinger, A. (2020) Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In *Lecture Notes in Computer Science*, Vol. 12279, 1–16.

Lundberg, S. M. – Erion, G. – Chen, H. – DeGrave, A. – Prutkin, J. M. – Nair, B. – katz, R. – Himmelfarb, J. – Lee, S-I. (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, Vol. 2(1), 56–67.

Mardani, A. – Nilashi, M. – Zavadskas, E. K. – Awang, S. R. – Zare, H. – Jamal, N. M. (2018). Decision Making Methods Based on Fuzzy Aggregation Operators: Three Decades Review from 1986 to 2017. *International Journal of Information Technology and Decision Making*, Vol. 17(2), 391–466.

Marino, D. L. – Wickramasinghe, C. S. – Manic, M. (2018). An Adversarial Approach for Explainable AI in Intrusion Detection Systems. In *Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society*, 3237–3243, Washington, DC., United States, 2018.

McGuinness, D. L. – Glass, A. – Wolverton, M. – Da Silva, P. P. (2007) A Categorization of Explanation Questions for Task Processing Systems. In *Proceedings of the 2007 AAAI Workshop on Explanation-Aware Computing (ExaCt 2007)*, 42–48, Vancouver, Canada, 2007.

Miller, T. (2019) Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, Vol. 267, 1–38.

Ming, Y. – Xu, P. – Cheng, F. – Qu, H. – Ren, L. (2019) ProtoSteer: Steering Deep Sequence Model with Prototypes. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 26(1), 238–248.

Moher, D. – Liberati, A. – Tetzlaff, J. – Altman, D. G. (2009) Preferred Reporting Items for Systematic Reviews and Meta-analyses: the PRISMA Statement. *Physical Therapy*, Vol. 89(9), 873–880.

Myers, B. A. – Weitzman, D. A. – Ko, A. J. – Chau, D. H. (2006). Answering Why and Why Not Questions in User Interfaces. In *Proceedings of the 2006 SIGCHI Conference on Human Factors in Computing Systems*, 397–406, Montréal, Canada, 2006.

Narayanan, M. – Chen, E. – He, J. – Kim, B. – Gershman, S. – Doshi-Velez, F. (2018) How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint*, arXiv:1802.00682.

Ngo, T. – Kunkel, J. – Ziegler, J. (2020) Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 183–191, Genoa, Italy, 2019.

Oh, C. – Song, J. – Choi, J. – Kim, S. – Lee, S. – Suh, B. (2018) I Lead, You Help but Only with Enough Details: Understanding the User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13, Montréal, Canada, 2018.

Olszewska, J. I. (2019). Designing Transparent and Autonomous Intelligent Vision Systems. In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence ICAART (2019)*, 850–856, Prague, Czech Republic, 2019.

Preece, A. (2018) Asking 'Why' in AI: Explainability of Intelligent Systems – Perspectives and Challenges. *Intelligent Systems in Accounting, Finance and Management*, Vol. 25(2), 63–72.

Prensky, M. (2001) Digital Natives, Digital Immigrants Part 2: Do They Really Think Differently? *On the Horizon*, Vol. 9(6), 1–6.

PRISMA (2021) PRISMA Endorsers. <http://www.prisma-statement.org/Endorsement/PRISMAEndorsers>, retrieved 11.3.2021.

Putnam, V. – Conati, C. (2019) Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS). In *Joint Proceedings of the ACM IUI 2019 Workshops*, 1–7, Los Angeles, United States, 2019.

Ribeiro, M. T. – Singh, S. – Guestrin, C. (2016). " Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144, San Francisco, United States, 2016.

Samek, W. – Wiegand, T. – Müller, K. R. (2017) Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries*, Special Issue No. 1 (2017), 1–10.

Samih, A. – Adadi, A. – Berrada, M. (2019). Towards a Knowledge Based Explainable Recommender Systems. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, 1–5, Tangier-Tetuan, Morocco, 2019.

Satalkina, L. – Steiner, G. (2020). Digital Entrepreneurship and its Role in Innovation Systems: A Systematic Literature Review as a Basis for Future Research Avenues for Sustainable Transitions. *Sustainability*, Vol. 12(7), 1–27.

Schaaf, N. – Huber, M., – Maucher, J. (2019). Enhancing Decision Tree Based Interpretation of Deep Neural Networks Through l1-orthogonal Regularization. In *Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA 2019)*, 42–49, Boca Raton, United States, 2019.

Schoenborn, J. – Althoff, K. (2019) Recent Trends in XAI: A Broad Overview on Current Approaches, Methodologies and Interactions. In *Proceedings of the ICCBR 2019 Workshops*, 51–60, Otzenhausen, Germany, 2019.

Schrills T. – Franke T. (2020) Color for Characters – Effects of Visual Explanations of AI on Trust and Observability. In: *Artificial Intelligence in HCI – HCII 2020*, eds. Degen, H. – Reinerman-Jones, L., *Lecture Notes in Computer Science*, vol. 12217. Springer, Cham.

Stumpf, S. (2019) Horses for Courses: Making the Case for Persuasive Engagement in Smart Systems. In *Joint Proceedings of the ACM IUI 2019 Workshops*, 1–6, Los Angeles, United States, 2019.

Thomas, J., – Haertling, T. (2020). AIBx, Artificial Intelligence Model to Risk Stratify Thyroid Nodules. *Thyroid*, Vol. 30(6), 878–884.

University of Turku (2013) Ethical Guidelines for Learning. <https://www.utu.fi/en/ethical-guidelines-for-learning>, retrieved 2.3.2021.

van den Berg, M. – Kuiper, O. (2020). XAI in the Financial Sector. Whitepaper, Hogeschool Utrecht, 2020.

van der Waa, J. – Schoonderwoerd, T. – van Diggelen, J. – Neerincx, M. (2020). Interpretable Confidence Measures for Decision Support Systems. *International Journal of Human-Computer Studies*, Vol. 144, 1–11.

van Mulken, S. – Andre, E. – Müller, J. (1998) The Persona Effect: How Substantial is it?. In: *People and Computers XIII*, eds. Johnson H. – Nigay L. – Roast C., 53–66. Springer, London.

Wang, D. – Yang, Q. – Abdul, A. – Lim, B. Y. (2019) Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15, Glasgow, United Kingdom, 2019.

Webster, J. – Watson, R. T. (2002) Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, Vol. 26(2), 13–23.

Weitz, K. – Schiller, D. – Schlagowski, R. – Huber, T. – André, E. (2019) "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9, Paris, France, 2019.

Weitz, K. – Schiller, D. – Schlagowski, R. – Huber, T. – André, E. (2019) "Let Me Explain!": Exploring the Potential of Virtual Agents in Explainable AI Interaction Design. *Journal on Multimodal User Interfaces*, 1-12.

Wieringa, M. (2020). What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT * 2020)*, 1–18, Barcelona, Spain, 2020.

Wohlin, C. (2014) Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10, London, United Kingdom, 2014.

Xie, Y. – Gao, G. – Chen, X. A. (2019) Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis. In *Joint Proceedings of the ACM IUI 2019 Workshops*, 1–8, Los Angeles, United States, 2019.

Yin, M. – Wortman Vaughan, J. – Wallach, H. (2019) Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12, Glasgow, United Kingdom, 2019.

Yu, K. – Berkovsky, S. – Conway, D. – Taib, R. – Zhou, J. – Chen, F. (2016) Trust and Reliance Based on System Accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, 223–227, Halifax, Canada, 2016.

Zhu, J. – Liapis, A. – Risi, S. – Bidarra, R. – Youngblood, G. M. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation.

In *Proceedings of the 2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8, Maastricht, The Netherlands, 2018.

APPENDICES

Appendix 1. Articles included in the sample

Authors	Title
Binns et al. (2018)	'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions
Brennen (2020)	What Do People Really Want When They Say They Want "Explainable AI"? We Asked 60 Stakeholders.
Broekens et al. (2010)	Do You Get It? User-Evaluated Explainable BDI Agents
Bussone et al. (2015)	The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems
Cai et al. (2019)	"Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human–AI Collaborative Decision-Making
Chazette and Schneider (2020)	Explainability as a Non-Functional Requirement: Challenges and Recommendations
Cheng et al. (2019)	Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders
Cirqueira et al. (2020)	Scenario-Based Requirements Elicitation for User-Centric Explainable AI
Cramer et al. (2008)	The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender
Dodge et al. (2019)	Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment
Ehsan et al. (2019)	Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions
Eiband et al. (2018)	Bringing Transparency Design into Practice
Eslami et al. (2018)	Communicating Algorithmic Process in Online Behavioral Advertising
Hohman et al. (2019)	Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models
Holstein et al. (2019)	Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?
Hong et al. (2020)	Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs
Liao et al. (2020)	Questioning the AI: Informing Design Practices for Explainable AI User Experiences

Lim and Dey 2009	Assessing Demand for Intelligibility in Context-Aware Applications
Lim et al. (2009)	Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems
Ngo et al. (2020)	Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study
Oh et al. (2018)	I Lead, You Help But Only with Enough Details: Understanding the User Experience of Co-Creation with Artificial Intelligence
Putnam & Conati (2019)	Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS)
Schrills & Franke (2020)	Color for Characters - Effects of Visual Explanations of AI on Trust and Observability
van der Waa et al. (2020)	Interpretable confidence measures for decision support systems
Wang et al. (2019)	Designing Theory-Driven User-Centric Explainable AI
Weitz et al. (2019)	"Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design
Weitz et al. (2020)	“Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design
Xie et al. (2019)	Outlining the Design Space of Explainable Intelligent Systems for Medical Diagnosis
Yin et al. (2019)	Understanding the Effect of Accuracy on Trust in Machine Learning Models

Appendix 2. Overview of the stakeholder groups and themes discussed (added with articles)