# How to explain AI systems to end users: a systematic literature review and research agenda

Samuli Laato
*Department of Computing, University of Turku, Turku, Finland and Gamification Group, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland*
Miika Tiainen
*Turku School of Economics, University of Turku, Turku, Finland*
A.K.M. Najmul Islam
*Turun Yliopisto, Tampere, Finland and LUT University, Lappeenranta, Finland, and*
Matti Mäntymäki
*Turun Kauppakorkeakoulu, Turku, Finland*

Received 26 August 2021
Revised 26 November 2021
10 January 2022
3 April 2022
Accepted 3 April 2022

## Abstract

**Purpose** – Inscrutable machine learning (ML) models are part of increasingly many information systems. Understanding how these models behave, and what their output is based on, is a challenge for developers let alone non-technical end users.

**Design/methodology/approach** – The authors investigate how AI systems and their decisions ought to be explained for end users through a systematic literature review.

**Findings** – The authors' synthesis of the literature suggests that AI system communication for end users has five high-level goals: (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) fairness. The authors identified several design recommendations, such as offering personalized and on-demand explanations and focusing on the explainability of key functionalities instead of aiming to explain the whole system. There exists multiple trade-offs in AI system explanations, and there is no single best solution that fits all cases.

**Research limitations/implications** – Based on the synthesis, the authors provide a design framework for explaining AI systems to end users. The study contributes to the work on AI governance by suggesting guidelines on how to make AI systems more understandable, fair, trustworthy, controllable and transparent.

**Originality/value** – This literature review brings together the literature on AI system communication and explainable AI (XAI) for end users. Building on previous academic literature on the topic, it provides synthesized insights, design recommendations and future research agenda.

**Keywords** Explainable AI, Explanatory AI, XAI, Machine learning, Human–computer interaction, End users, Literature review, Systematic literature review

**Paper type** Research paper



The initial literature search upon which this article develops was done for the following Master's thesis published at the University of Turku:

Tiainen, M., (2021), To whom to explain and what?: Systematic literature review on empirical studies on Explainable Artificial Intelligence (XAI), available at: https://www.utupub.fi/handle/10024/151554, accessed April 2, 2022.

Internet Research
Vol. 32 No. 7, 2022
pp. 1-31
Emerald Publishing Limited
1066-2243
DOI 10.1108/INTR-08-2021-0600

## 1. Introduction

Artificial intelligence (AI) systems are becoming increasingly complex (Karamitsos *et al.*, 2020; von Eschenbach, 2021). This trend can be attributed to advances in machine learning (ML) model technology, that has advanced towards better predictive power, but as a consequence, the models have become inscrutable and more difficult to explain (Brennen, 2020; Došilović *et al.*, 2018; von Eschenbach, 2021). Simultaneously, AI functionalities are being integrated as part of a growing range of information systems (Hornung and Smolnik, 2021; Rana *et al.*, 2021; Tarafdar *et al.*, 2019) and used to support critical decision-making. For example, ML approaches have been used to combat the COVID-19 pandemic through patient outcome prediction, risk assessment and predicting the disease spreading (Dogan *et al.*, 2021), and are an integral component of recommendation systems that curate social media feeds and e-commerce (Batmaz *et al.*, 2019). To reinforce public trust in AI-driven and AI-supported decision making, and to mitigate prejudices (Zarifis *et al.*, 2020) it is pivotal to ensure the explainability of AI-made decisions to the end users of these systems (European Commission, 2020).

The increased deployment of AI, particularly in high-risk and critical application areas such as military (Dawes, 2021) and healthcare (Smith, 2021), has spurred a public debate on the risks and unintended negative consequences of ill-governed black-box algorithms (Jobin *et al.*, 2019; Liang *et al.*, 2021; Shneiderman, 2020). The potential missteps of ML system decisions, and misinterpretations of ML model output due to lack of understanding, have potentially grave consequences (Rana *et al.*, 2021). Simultaneously, AI systems are increasingly being used by individuals with non-technical backgrounds (Liang *et al.*, 2021) such as medical doctors and clinicians (Bussone *et al.*, 2015; Cai *et al.*, 2019) and lawyers (Dodge *et al.*, 2019). This has created the need to come up with AI system explanations and communication aimed for end users to ensure trust through transparency [1] (European Commission, 2020). However, explaining AI systems and communicating about them to end users are not straightforward tasks (Weitz *et al.*, 2021).

Previous research has delineated several potential barriers to the explainability of AI systems, including technical challenges (Anjomshoae *et al.*, 2019), limitations of human logic (Asatiani *et al.*, 2020, 2021) and even intentional secrecy (Burrell, 2016). However, even with full explanations, the issue of how to communicate about the AI system to end users remains a challenge (Brennen, 2020). For example, for end users, a completely transparent, full explanation may not always be the most useful one (Lim *et al.*, 2009; Broekens *et al.*, 2010). The field of explainable AI (XAI) research seeks to bring clarity to how specific ML models work. According to Arrieta *et al.* (2020), XAI can be defined as follows: "*Given an audience, an XAI is one that produces details or reasons [regarding a ML model] to make its functioning clear or easy to understand.*" Hence, XAI is crucial to ensure that AI systems produce sufficient information regarding their operation that allows explanations to be given about the system to their users. Communicating AI explanations to end users represents a key challenge for AI system design and an important area of study for XAI research. This calls for review studies providing evidence-based insights about end users' explainability needs and preferences, as well as synthesizing work to uncover best practices for designing end user AI communication suggested in previous XAI research.

While prior XAI literature features a few systematic literature reviews (SLRs), no SLR has specifically focused on end users as an audience or users of XAI. Arrieta *et al.* (2020) and Anjomshoae *et al.* (2019) investigated technical solutions for XAI. Anjomshoae *et al.* (2019) discovered that the literature featured only relatively few studies focusing on XAI and AI system explanations for end users. The SLR conducted by Antoniadi *et al.* (2021), in turn, elaborated on XAI for clinical decision support systems and their users, with also a primary focus on technical solutions. In this study, we depart from the technical XAI literature (Anjomshoae *et al.*, 2019; Antoniadi *et al.*, 2021; Arrieta *et al.*, 2020) and focus on studies on AI system end user communication. In doing so, we answer the following research questions:

*RQ1.* What are the goals and objectives of AI system explanations for end users?

*RQ2.* What recommendations does the extant literature suggest for designing explanations for AI systems that facilitate positive outcomes?

*RQ3.* What future research directions arise from the extant literature?

Through answering the research questions we make three contributions. First, we answer the recent calls to study AI system end users and XAI from the HCI perspective (Brennen, 2020; Weitz *et al.*, 2019b). Second, we summarize and synthesize the findings of extant empirical studies on five objectives of XAI (Meske *et al.*, 2022) for end users. Third, we provide an agenda for future research in this field. The rest of this study is structured as follows. In the background section, we look at previous studies on technical XAI solutions to determine what kinds of explanations are possible, followed by the identification of the stakeholders of XAI to determine who the end users of XAI are and what the search keywords are for the SLR work. Next, we describe the methods and data collection process for the SLR, followed by the findings concerning the three research questions. We conclude the paper with a discussion of the results, theoretical and practical implications, limitations and future work.

## 2. Background

### *2.1 Technical XAI solutions*

XAI can be considered the starting point of AI system explanations for end users. Arrieta *et al.* (2020) classified model agnostic post-hoc XAI techniques into four categories: (1) explanation by simplification, where the AI system is explained by simplifying it either through architecture modification or other means; (2) feature relevance explanation, where the relevance of the features that contribute to a specific model decision are highlighted; (3) local explanation, where parts of the larger model are explained individually, and (4) visual explanation, which aims to provide visual support such as heat maps for machine vision algorithms that help understand what factors the model prediction was based on (Arrieta *et al.*, 2020). These categories are not mutually exclusive, and, for example, methods such as local interpretable model-agnostic explanations (LIME) belong to both explanation by simplification and local explanation categories (Ribeiro *et al.*, 2016). Other widely used XAI tools include SHAP [2] and its derivatives (Ribeiro *et al.*, 2016), which aim to provide various visualizations that can demystify the inner workings of ML models and visualize the process that ultimately generates the models' predictions. In practice, this can be executed through, for example, heat maps for computer vision algorithms and graphs displaying which factors inside the model had the biggest impact on the final decision (Parsa *et al.*, 2020). Moreover, solutions such as Google Model Cards aim to present a clear, transparent report of ML models (Mitchell *et al.*, 2019). The Model Card is not a technical XAI approach as such; rather, it delivers knowledge of what data was used to train and test the model. Hence, it also can include XAI reports and visualizations (Mitchell *et al.*, 2019).

Another approach to increasing model explainability is to design interpretable ML systems from the beginning (Evans *et al.*, 2021). These types of models can be regarded as transparent because they are explainable by themselves. Examples of such systems include rule-based systems, Bayesian models and decision trees (Arrieta *et al.*, 2020). Recently, researchers have managed to create transparent unsupervised learning models, for example, via a neural-symbolic computing approach (Evans *et al.*, 2021). Such approaches also yield novel opportunities for AI system communication for end users. Overall, XAI technology is constantly advancing, and the technical solutions (ML approach and the explainability) have enormous influence on what can be explained from a certain ML model.

### *2.2 XAI stakeholders*

Table 1 presents the five key stakeholder groups for XAI which are commonly discussed in the extant literature (Meske *et al.*, 2022; Arrieta *et al.*, 2020), and a rationale for each group.

| XAI stakeholder group | Explanation |
|---|---|
| Users affected by model decisions | As AI systems are widely implemented across services, people are constantly directly and indirectly affected by decisions made by various models in various contexts |
| Individuals using AI systems | An increasing number of tools and services include AI components. Examples are numerous, from online recommendation systems to anomaly detection solutions trying to block spam email |
| Managers and executive board members | In business firms, upper executive management has oversight into the AI systems used in their company |
| Regulatory entities | Various regulatory entities such as the European Union and individual governmental bodies are interested in controlling and legislating AI systems to protect citizens from potential harm that immature AI systems could cause |
| Developers such as data scientists and system engineers | Perhaps the most obvious target audience for XAI are the developers who create the AI models. They are responsible for ensuring that the models work effectively and in a desired fashion |

**Source(s):** Based on Meske *et al.* (2022)

**Table 1.**
The five stakeholder groups of XAI

Regarding the end users of AI systems, the literature distinguishes between individuals voluntarily using AI systems and individuals affected by decisions made by AI systems (Meske *et al.*, 2022). In addition, there are two stakeholder groups overseeing AI systems from different perspectives. Regulatory entities ensure that AI systems comply with laws and regulations, while managers and executive board members make sure AI systems serve their purpose in the overall business landscape. Finally, there are AI system developers who are considered a stakeholder group of their own (Arrieta *et al.*, 2020; Meske *et al.*, 2022). Against this backdrop, AI system end users entail both the people who use AI systems and the people who are influenced by the AI system's decision-making. This duality is particularly exemplary in today's AI-driven consumer services. While AI end users are doing a Google search or browsing Netflix, they are constantly both using an AI system and being influenced by its decision making (Ngo *et al.*, 2020).

Meske *et al.* (2022) suggest that there are five general reasons for implementing XAI: (1) evaluating the AI; referring to forming an idea of how well the AI system performs, (2) justifying the AI, referring to ensuring the system works in a correct and fair manner, (3) learning from the AI, referring to increasing understanding of the system, (4) improving the AI, referring to the capability to make the AI system better, and connected to all these (5) managing the AI, that is, ensuring the AI system stays under control and operates as intended. Meske *et al.* (2022) further argued that these objectives might differ between XAI stakeholder groups. When implementing XAI and transparent AI in practice, it is important to remember the audience (Parsa *et al.*, 2020; Ribeiro *et al.*, 2016). As an example, laypeople on average do not possess the same technical abilities and knowledge of ML systems as data scientists or AI auditors (Dodge *et al.*, 2019; van der Waa *et al.*, 2021; Weitz *et al.*, 2019a). Thus, with regards to the end users, it is important to elucidate the goals and drivers of XAI and the communication of explanations.

## 3. Methodology

### *3.1 Literature search*

To systematically identify studies on XAI from the HCI perspective, we conducted a preliminary subjective examination of the topic and identified relevant keywords and terminology. We discovered various terms that have been used in XAI research to describe

the concept. These include AI *interpretability*, *transparency* and *understandability*. XAI research is also connected to the topics of AI accountability, responsibility and governance. The lack of consistent terminology has recently been discussed by, for example Brennen (2020), who identified through stakeholder interviews that practitioners are, in fact, using up to 20 synonyms for XAI. Drawing from the interviews conducted by Brennen (2020), we summarized a list of alternative terms for XAI that are relevant in our SLR. These included *explanatory AI*, *transparent AI*, *interpretable AI* and *accountable AI*. Concerning HCI, we decided not to include specific keywords, but rather filter out studies at a later stage because we noticed that HCI is not a keyword included in many of the studies that seemed to fit the scope of our work. Based on this work, we formulated search strings, which are available in Appendix 1.

In conducting the literature search, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher *et al.*, 2009). We looked up literature from two popular meta-level research databases: Scopus and Web of Science Core Collection. Scopus is known for indexing databases that are relevant to computer science, such as ACM, DBLP Computer Science Bibliography, IEEExplore and SpringerLink (Morschheuser *et al.*, 2017). Supplementing the search with Web of Science Core Collection increases the robustness of the results. According to Kitchenham and Brereton (2013), screening the references of selected studies can help discover studies that were missed during the initial search. This is particularly relevant in our case since we focused on XAI specific keywords and omitted search terms related to AI system communication. Thus, we also performed backward snowballing (Wohlin, 2014), where we went through the references of our resulting sample of papers, identified potential studies connected to the research topic, examined all their references, and repeated this process until no new studies emerged (Wohlin, 2014).

Both Scopus and Web of Science Core Collection were searched in October 2020. The bibliographic information of the studies was downloaded in .csv format and subsequently combined. The search in Scopus resulted in 723 articles, and the Web of Science Core Collection resulted in 325 articles. Upon combining articles from both databases and removing duplicates, 808 articles remained.

### *3.2 Inclusion and exclusion criteria*

The inclusion and exclusion criteria for the study item processing are displayed in Table 2. These criteria were used in the identified articles in two stages. In the first stage, only the titles and abstracts of the studies were read. As the full paper articles were not evaluated at this stage, we wanted to be broad with the inclusion criteria to avoid false negatives. Accordingly, if any of the criteria in Table 2 were met, the study was included in the second phase, where full texts were assessed for eligibility.

Out of the 808 initial articles, 620 were excluded based on the inclusion and exclusion criteria specified in Table 2. Examples included studies not related to AI, studies in math and chemistry where "XAI" was part of a formula, studies conducted near the Mozambican city Xai-Xai, and studies involving the indigenous Xai'xai people from Canada. Furthermore,

| Inclusion criteria | Exclusion criteria |
|---|---|
| 1. The abstract of the study indicates the study is focused on XAI aimed at end users or ML-based system explanations for end users<br>2. The research specifically approaches the issue from an HCI perspective<br>3. The research is empirical | 1. Editorials, opinion papers or other non-peer-reviewed work<br>2. Studies in languages other than English |

**Table 2.** Inclusion and exclusion criteria in the screening phase

we excluded non-empirical work (based on criterion #3) and technical AI studies that were clearly not related to the end users of AI systems. Unclear and borderline cases were included at this stage to avoid false negatives. Thus, after screening for the abstract and title, we were left with 188 studies. This process was conservative and straightforward and conducted by one of the authors.

In the second stage, we assessed the full texts of the studies to determine whether they concerned XAI or AI explanations for end users. At this stage, if it was not clear whether a study should be included, we discussed it between the authors until a decision was reached. Based on these discussions we omitted, for example, studies that focused on the feasibility of a specific XAI solution (Kuwajima *et al.*, 2019; Ming *et al.*, 2019) and non-empirical studies on XAI stakeholders (e.g. Zhu *et al.*, 2018). Subsequently, we were left with 19 articles, which we proceeded to engage in backward chaining (Wohlin, 2014). Accordingly, we screened the reference lists of all 19 articles. In case an article seemed potentially related to the research topic, we looked it up and read the abstract. If the article still seemed relevant, we read the full text. Applying the same inclusion and exclusion criteria as before, we proceeded to comb all the references. If the article was included, we also read its references, repeating the process (Wohlin, 2014). Through this procedure, we identified six additional articles, resulting in the final number of 25 articles to be included in the synthesis. The entire data search process is displayed in Figure 1.

### *3.3 Data extraction and analysis*

With the final sample of studies ($n = 25$), we agreed on a specific set of information that we systematically extracted to answer the three research questions. The fetched descriptive information was as follows: (1) publication venue, (2) publication year, (3) study approach, (4) methodology and (5) sample. Subsequently, we extracted information on the studied end user groups, in order to obtain knowledge on possible differences among XAI needs between the groups. To answer the research question RQ1*:* "*What are the goals and objectives of AI system explanations for end users?*", we extracted the studied outcomes of AI communication. We went through the empirical studies and assigned codes for each measured or investigated goal and objective. This approach was similar to open coding (Strauss and Corbin, 1998). We then conducted a round of axial coding where we sought to combine similar codes together to form thematic clusters. For example, intelligibility, comprehensiveness and understandability were combined into the same theme; as were justice and fairness. In the end, five thematic clusters merged as the objectives for AI system explanations for end users. When discussing these, we returned to the codes and looked at the results and discussion surrounding each theme from the research papers.

For RQ2: *What recommendations does the extant literature suggest for designing explanations for AI systems that facilitate positive outcomes?*, we extracted each unique design recommendation that appeared either explicitly or implicitly in the studies. Similarly to the analysis process for answering RQ1, we coded the design recommendations from the studies using open coding (Strauss and Corbin, 1998) and then combined similar codes together. We also extracted information regarding the context in which the given recommendations apply and organized the recommendations into general, what to explain, how to explain and when to explain. Finally, for RQ3: "*What future research directions arise from the extant literature?*", we searched for the future research directions presented in the studies. We extracted each explicitly stated research direction, but also conducted a meta-level synthesis on the research directions that arise from the extant literature on XAI for end users as a whole.

### *3.4 Descriptive data of reviewed studies*

With respect to the publication venues, the most common outlet was the Proceedings of the *CHI Conference on Human Factors in Computing Systems,* with six studies. The second most

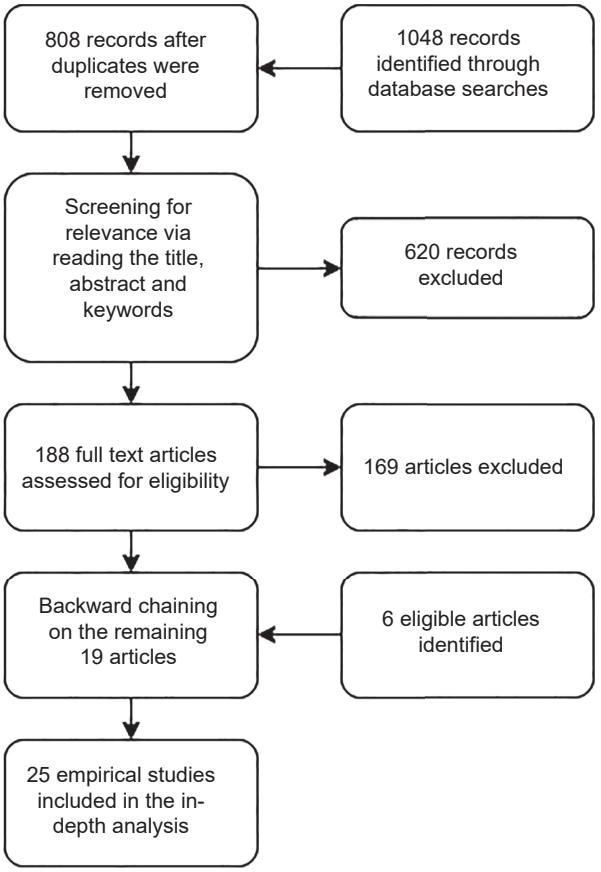


**Figure 1.**
The literature search and screening process

popular outlet was the *Proceedings of the International Conference on Intelligent User Interfaces* (three studies). Of the identified studies, most were published in 2018–2020 ($n$ = 20), with some preliminary work already taking place in 2008–2010. The publication years of the studies are displayed in Figure 2.

All studies featured human participants in some form or another, which was expected, as we specifically looked for empirical studies on XAI in the field of HCI. Five studies drew participants from the USA (Brennen, 2020; Cai *et al.*, 2019; Eslami *et al.*, 2018; Putnam and Conati, 2019; Xie *et al.*, 2019), two from the UK (Binns *et al.*, 2018; Bussone *et al.*, 2015), and two from the Netherlands (Broekens *et al.*, 2010; Cramer *et al.*, 2008). Other countries from which participants were selected included South Korea (Oh *et al.*, 2018), Germany and Brazil (Chazette and Schneider, 2020), and Austria (Cirqueira *et al.*, 2020). In addition, not included in the countries listed above were studies that sourced their participants from online crowdsourcing websites such as MTurk (Cheng *et al.*, 2019; Dodge *et al.*, 2019; Lim and Dey, 2009; van der Waa *et al.*, 2020; Yin *et al.*, 2019), TurkPrime (Ehsan *et al.*, 2019) and Prolific Academic (Binns *et al.*, 2018). Various studies did not specify where their participants were recruited, but instead had a heavy focus on the AI system itself, and its evaluation was only secondary. All the studies including information regarding their approach and methods are available in Table A1.

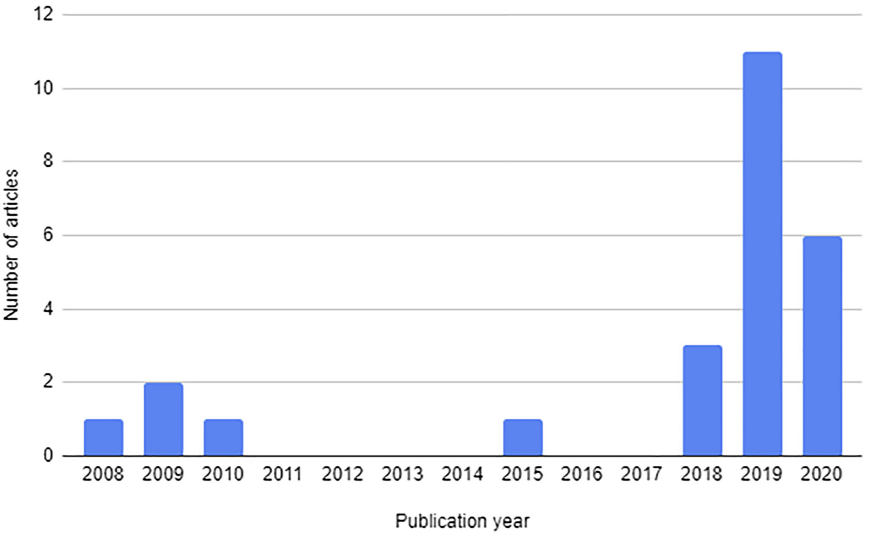


**Figure 2.**
Papers included in the study by publication year

## 4. Results

### *4.1 End users and application contexts of AI systems*

We looked at the application contexts in which academic studies have observed XAI needs of end users. Most of the studies included in the review focused on a specific user group, however, some included several scenarios (e.g. Binns *et al.*, 2018) and some looked at AI systems generally without explicitly connecting to a specific context (see Chazette and Schneider, 2020; Lim and Dey, 2009; Lim *et al.*, 2009; Schrills and Franke, 2020; van der Waa *et al.*, 2020). While the contexts were various, the ML approaches were sometimes similar. For example, outcome prediction systems were used to predict speed dating scenarios (Yin *et al.*, 2019) and outcomes of criminal trials (Dodge *et al.*, 2019).

The application contexts in the studies are described in Table 3. In addition, there were studies which did not specify a context but looked at AI systems generally (e.g. Chazette and Schneider, 2020; Lim and Dey, 2009; Schrills and Franke, 2020). By identifying the application contexts, we also identified end users who are likely to benefit from AI system explanations in that context. We notice that XAI and end user communications needs to be aimed at least towards the following stakeholder groups: laypeople, doctors, other medical professionals, clerks, tellers, actuaries, sales personnel, human resources personnel, administrative staff, management staff, airline employees, security specialists, IT personnel, financial crime specialists, judges, jury members, defendants, prosecutors, attorneys and employees working for technology providers. This large but not exhaustive list corresponds with the reported ubiquitous proliferation of ML across IS (Collins *et al.*, 2021; Laato *et al.*, 2021, 2022).

Laypeople could be pinpointed as a key end user group in a multitude of studies (e.g. Eslami *et al.*, 2018; Schrills and Franke, 2020; van der Waa *et al.*, 2020; Yin *et al.*, 2019), but Table 3 highlights that there are also various professional groups and even artists (Oh *et al.*, 2018) who are dealing with AI systems in a way where the users are likely to benefit from XAI solutions. These observations underscore how XAI tools and systems should be built not only for developers (Arrieta *et al.*, 2020), and not only for end users generally (Meske *et al.*, 2022), but to various professionals and expert groups across industry sectors who, depending on even more specific application contexts within their field, may wish for various kinds of explanations concerning the system.

| Application context | End users | Description | Studies |
|---|---|---|---|
| Medical decision making | Doctors, other medical professionals | Medical professionals use AI systems to assist in critical decision making such as detecting tumors from CT scans | Bussone *et al.* (2015), Cai *et al.* (2019), Wang *et al.* (2019), Xie *et al.* (2019) |
| Loans and finance | Laypeople, clerks | Banks can automate loan application processes with ML systems | Binns *et al.* (2018) |
| Insurance | Laypeople, actuaries, sales personnel | Insurance pricing is a complex endeavor due to the multitude of data sources influencing the optimal price. AI systems can help form this price | Binns *et al.* (2018) |
| Explanation agents and automatic rationale generation for daily activities | Laypeople | Explanation agents provide reasoning and justification for actions across a wide range of fields and topics including cooking and gaming | Broekens *et al.* (2010), Ehsan *et al.* (2019), Weitz *et al.* (2019a, b) |
| Evaluation of applications (work, university admission, promotion) | Human resources, company management, administrative staff, laypeople | Larger companies can have internal AI tools for evaluating worker performance, which can be used in decision making when employees apply for promotions. Similar processes can be used for university admissions or going through work applications | Binns *et al.* (2018), Cheng *et al.* (2019) |
| Re-routing passengers for overbooked flights | Airline personnel | Airlines sometimes overbook flights and have to re-route passengers. AI systems can help design new routes for passengers | Binns *et al.* (2018) |
| Fraud detection | Security specialists, IT personnel, financial crime specialists | Fraud detection specialists can use anomaly detection ML models and other tools to discover unusual and hence suspicious events | Binns *et al.* (2018), Cirqueira *et al.* (2020) |
| Criminal trials | judges, jury, the defendant, prosecutors, attorneys | AI systems can be used in court to provide decisions, or decision support for the involved stakeholders | Dodge *et al.* (2019) |
| Online advertising | Laypeople, advertising agencies, technology providers | Online advertisements are almost ubiquitously based on ML systems which determine based on profiling data which ads to show and to whom | Eslami *et al.* (2018) |
| Speed dating | Relationship advisors, laypeople | A prediction system can be put in place to estimate the outcome of speed dating scenarios | Yin *et al.* (2019) |

(*continued*)

**Table 3.**
Application context of the XAI in HCI studies

| Application context | End users | Description | Studies |
|---|---|---|---|
| E-commerce, social media, tutoring systems | Laypeople, students, trainee employees | Recommender systems can be based on the user's own previous history or data collected from other similar users and their preferences. These systems are in operation all over the Internet, for example, in e-commerce and social media | Cramer *et al.* (2008), Eiband *et al.* (2018), Ngo *et al.* (2020), Putnam and Conati (2019) |
| Human-AI Co-creation | Laypeople, artists, other employees in creative fields | AI systems can automate parts of creative processes. For example, in music composing AI can propose all kinds of melodies and the job of the composer is to pick those that are relevant | Oh *et al.* (2018) |

**Table 3.**

*4.2 The objectives and goals of AI communication for end users*

Based on the extraction of the key goals of XAI from the empirical studies, we identified five key objectives or goals for explaining AI systems for end users. These were the increasing of (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) the fairness of the system. Studies also discussed general goals not particularly related to the ML system, itself, such as usability, ease of use and satisfaction (Oh *et al.*, 2018). Furthermore, studies have approached these goals from two main intertwined perspectives: features of the system and perceptions of the end users. In practice, the features of the system were obtained via observing the perceptions of the end users; thus, the two are discussed together in this section. The identified central objectives of XAI and AI system communication for end users are displayed in Table 4. For each objective, we identified the most popular term but also included synonyms and other words or concepts that were indistinguishable from the main objective.

*4.2.1 Understandability.* When discussing how end users understand an AI system, four terms were used. In addition to *understandability*, the three other terms were *interpretability* (Chazette and Schneider, 2020), *comprehensibility* (Oh *et al.*, 2018) and *intelligibility* (Ehsan *et al.*, 2019; Lim and Dey, 2009). Slightly depending on the definition and interpretation, these four terms all referred to how accurately end users could imagine the system's operation and its decisions. In addition, there was significant overlap in the studies between the understandability (1) of the system and (2) of the communication. For example, a few studies focused on the characteristics of the system (Lim and Dey, 2009; Lim *et al.*, 2009), and from there aimed to work toward how end users perceived it. By contrast, other studies approached understandability from the perspective of end users' perceptions, and here, communication about the system was highlighted (Cheng *et al.*, 2019; Cirqueira *et al.*, 2020; Dodge *et al.*, 2019; Ehsan *et al.*, 2019). The understanding of the system was also connected to the other identified themes. The transparency of the system, for example, enabled users to better understand it (Cramer *et al.*, 2008).

Among the various approaches toward understandability, there was a consensus that end users' technical knowledge, prior conceptions, and mental abilities must be considered when explaining AI systems to them (e.g. Chazette and Schneider, 2020; Dodge *et al.*, 2019; Ehsan *et al.*, 2019; Weitz *et al.*, 2019a; Wang *et al.*, 2019; Xie *et al.*, 2019). The better aligned the explanations were with the users' conceptions and mental models, the better they understood the explanations (Ehsan *et al.*, 2019; Ngo *et al.*, 2020). This was particularly relevant for AI systems whose users with little technical expertise have a particularly high likelihood of

| Objectives/goal | Definition (adopted from the given sources) | Sources |
|---|---|---|
| Understandability<br>*Other included codes*: Intelligibility, comprehensibility, interpretability | The degree to which end users are able to form an accurate mental model regarding how the AI system works | Chazette and Schneider (2020), Cheng *et al.* (2019), Cirqueira *et al.* (2020), Cramer *et al.* (2008), Ehsan *et al.* (2019), Eiband *et al.* (2018), Eslami *et al.* (2018), Lim *et al.* (2009), Lim and Dey (2009), Oh *et al.* (2018), Putnam and Conati (2019), van der Waa *et al.* (2020), Weitz *et al.* (2019a), Wang *et al.* (2019), Xie *et al.* (2019) |
| Trustworthiness<br>*Other included codes*: Trust | Refers to end users' perception about the truthfulness and honesty of the system, as well as beliefs that the system works as intended | Brennen (2020), Bussone *et al.* (2015), Cheng *et al.* (2019), Ehsan *et al.* (2019), Schrills and Franke (2020), Wang *et al.* (2019), Weitz *et al.* (2019a), Weitz *et al.* (2019b), Xie *et al.* (2019), Yin *et al.* (2019) |
| Transparency | The degree of information that is disclosed about the AI system. For example, high transparency systems disclose (almost) fully the system functioning from data to algorithms and parameters | Brennen (2020), Cai *et al.* (2019), Chazette and Schneider (2020), Cramer *et al.* (2008), Eiband *et al.* (2018), Ngo *et al.* (2020), Schrills and Franke (2020) |
| Controllability | End users' subjective sense of control over the AI system | Ngo *et al.* (2020), Oh *et al.* (2018), Wang *et al.* (2019) |
| Fairness<br>*Other included codes*: Justice | Refers to the subjective perception of whether the decisions or recommendations made by the AI system feel right and just | Binns *et al.* (2018), Dodge *et al.* (2019) |

**Table 4.**
XAI objectives

having misconceptions about how AI systems arrive at conclusions (Oh *et al.*, 2018; Xie *et al.*, 2019). Hence, XAI tools need to focus on delivering explanations that are intuitive, meaning visualizations and even metaphors to make the system more understandable for the end users (Schrills and Franke, 2020; Weitz *et al.*, 2019b).

A few studies focused specifically on the communication of AI systems, how to produce it, and how it is perceived by the end users. Studies have assessed both verbal (Eslami *et al.*, 2018; Ngo *et al.*, 2020) and nonverbal (Cheng *et al.*, 2019; Weitz *et al.*, 2019a) communication. Regarding verbal communication, oversimplified language and overly complicated language hinder end users' ability to understand a system (Eslami *et al.*, 2018). Providing overly complex and detailed explanations could cause information overload among the participants (Oh *et al.*, 2018), and according to Ehsan *et al.* (2019), contextual accuracy is more important than the length of the explanation. In summary, the understandability of AI systems relies on the system itself, verbal and non-verbal communication about it, and the end users' experiences and expertise.

*4.2.2 Trustworthiness.* Unlike understandability, trustworthiness is discussed primarily from the perspective of the end users, not the system. Yet, the term *trustworthiness* appeared ubiquitously in the selected studies referring to a characteristic of the AI system. However, the empirical studies focused on end users' trust in the system, which of course was influenced by how end users received knowledge about the system – meaning communication (Brennen, 2020; Bussone *et al.*, 2015; Cheng *et al.*, 2019; Ehsan *et al.*, 2019; Schrills and Franke, 2020; Yin *et al.*, 2019).

One of the interesting findings concerning trust in AI systems was that, in one study, XAI and clear explanation interfaces did not facilitate trust. They did, however, increase the understandability of a system (Cheng *et al.*, 2019). Even in a study where end users were shown that the AI system performs more accurately than the participant, their trust in the system did not increase (Yin *et al.*, 2019). By contrast, end users had more trust in explanations shown by virtual agents than, for example, only text- or voice-based explanations (Weitz *et al.*, 2019a, b). Furthermore, according to the findings of Schrills and Franke (2020), the relationship between explanation types and end users' trust in a system and its decisions is complex and not straightforward.

Four articles discussed medical professionals' perceived trust in XAI (Bussone *et al.*, 2015; Cai *et al.*, 2019; Xie *et al.*, 2019; Wang *et al.*, 2019). Here, the consensus was that medical professionals need in-depth explanations regarding why the system makes decisions (Bussone *et al.*, 2015; Cai *et al.*, 2019). Studies recommended XAI for medical professionals be as complete as possible and suggested systems to deliver the training data, source, and situational data, and other forms of external information to the system's users on demand (Xie *et al.*, 2019; Wang *et al.*, 2019). Thus, at least for medical professionals and other expert end users, XAI aiming to build trust should be transparent. Similar findings have appeared in studies with other types of end users (Cramer *et al.*, 2008; Ehsan *et al.*, 2019).

*4.2.3 Transparency.* As pointed out earlier, transparency was closely related to the discussion on trust, as the lack of transparency can adversely impact trust (Ehsan *et al.*, 2019) but also understandability, as perceived transparency is connected to how well users can understand content (Cramer *et al.*, 2008). While transparency could be measured objectively as a characteristic of the system, in the reviewed literature, transparency was primarily scrutinized from the viewpoint of end users (Brennen, 2020; Schrills and Franke, 2020; Ngo *et al.*, 2020; Eiband *et al.*, 2018). While some studies were conducted with AI systems created specifically for research purposes (e.g. Cai *et al.*, 2019), others investigated existing popular systems such as the Netflix content recommendation system (e.g. Ngo *et al.*, 2020). Regarding Netflix, participants in the study sample had inaccurate mental models of how it works. For example, some participants imagined the system would use much more information about them than it did in reality, while others did not realize the system would also use data from other Netflix users (Ngo *et al.*, 2020).

Compared to the rest of the identified objectives of XAI, transparency was seemingly straightforward, as it could objectively be defined as simply disclosing more information about the system for end users (Brennen, 2020; Cai *et al.*, 2019; Chazette and Schneider, 2020; Cramer *et al.*, 2008). However, the situation was not clear cut, as AI systems are not fully transparent even for the developers making them (Arrieta *et al.*, 2020), which brings the technical XAI perspective into the discussion. Besides what can be explained, the discussion on transparency includes the perspective of what information about the system is relevant for the end users (Eiband *et al.*, 2018).

*4.2.4 Controllability.* Altogether, three studies focused on end users' perceived sense of control over a system and, consequently, the system characteristic of "controllability" (Ngo *et al.*, 2020; Oh *et al.*, 2018; Wang *et al.*, 2019). In their model based on previous results of XAI research, Wang *et al.* (2019) postulated that one of the reasons people want explanations is to control and predict how the system behaves. Perceived control is, thus, an intrinsic need for system end users, which can be especially important if the system behaves in an unexpected or undesired manner (Wang *et al.*, 2019). While Wang *et al.* (2019) tested their framework with medical professionals, the other two studies focused on recommender systems (Ngo *et al.*, 2020) and a drawing tool involving AI (Oh *et al.*, 2018).

Ngo *et al.* (2020) focused on online recommender systems that use other people's data, the end user's own data, as well as potential other sources of data to find recommendations on what the user may like or what the system provider may want the user to click. The results indicate that, for end users to feel more in control of the system, the system's explanations

need to guide users to form mental models of the system that match its real technical implementation (Ngo *et al.*, 2020). Oh *et al.* (2018) created a drawing tool in which end users could draw images together with an AI system. Their results showed that end users wanted to oversee the drawing procedure and wanted the AI system to explain itself upon request (Oh *et al.*, 2018). Based on the findings of this study, interaction opportunities that enhance the sense of control for end users, such as the ability to command AI in various ways and the ability to choose when the AI system explains itself, are important for increasing the perceived controllability of the system (Oh *et al.*, 2018).

*4.2.5 Fairness.* Compared to the other goals of XAI reported above, the fairness cluster contained the smallest number of studies. Fairness and justice were discussed together and even appeared interchangeably (Binns *et al.*, 2018; Dodge *et al.*, 2019). The two articles (Binns *et al.*, 2018; Dodge *et al.*, 2019) focused particularly on end users' conceptions and perceptions regarding the fairness of AI systems. In both studies, participants viewed case-based explanations (i.e. explanations comparing the current case to previous cases) as least fair. Participants in Dodge *et al.* (2019) stated the following reasons: (1) case-based explanations do not provide adequate information about how an AI system arrives at a conclusion, (2) the number of cases provided in the experiment was considered too small, and (3) it is questionable whether one case can ever be considered identical to another.

In contrast, sensitivity-based explanations were ranked the fairest (Binns *et al.*, 2018; Dodge *et al.*, 2019). They were valued for their conciseness, understandability, and transparency when the decisions were non-controversial (Binns *et al.*, 2018; Dodge *et al.*, 2019). Interestingly, both case- and sensitivity-based explanations were local explanation styles (Arrieta *et al.*, 2020), as opposed to global explanations. It seems end users appreciate explanations that they understand, and conciseness and understandability are more important than an explanation's completeness. Finally, major individual differences exist, as Dodge *et al.* (2019) point out 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.*" Finally and interestingly, the studies discussing fairness pointed out that users will not trust the model or consider it fair regardless of improvements made to it if they consider the system's task type fundamentally unfit for algorithmic decision making (Binns *et al.*, 2018; Dodge *et al.*, 2019).

### *4.3 Design recommendations for explaining AI systems to end users*

Table 5 summarizes design recommendations for explaining AI system decisions, and communication about them, for end users presented in the reviewed studies. We have categorized the recommendations into four groups: general recommendations and recommendations addressing "when," "what" and "how" to explain. We identified 16 unique design recommendations. The heterogeneity of the recommendations stems from the differences in the research setups, study contexts, and research focuses of the studies (see Table A1), as well as the novelty and, thus, the formative stage of the research area.

The recommendations vary in specificity. While some are general, such as the one that guides designers to consider the context in which they provide AI system explanations for end users, others concern a specific aspect of the explanation design. Recommendation 2, which suggests providing explanations on demand, is the only example of "when" to explain; however, it was posited by several studies (Chazette and Schneider, 2020; Cramer *et al.*, 2008; Lim *et al.*, 2009; Lim and Dey, 2009; Oh *et al.*, 2018). Recommendation #8, the most often mentioned, suggests strengthening users' curiosity in the system (Oh *et al.*, 2018; Putnam and Conati, 2019). This along with recommendations #3–13 belong to the "how" category. Recommendation #15, which suggests users may want explanations for negative or less favorable AI decisions (Putnam and Conati, 2019), is an example of "what" to explain. Recommendations #14–16 fall into this category.

| Recommendation categories | Design recommendation | Reasoning | Sources |
|---|---|---|---|
| General | 1. Context is everything – There is no one-size-fits-all type of solution | What to explain is dependent on several factors including what kind of AI system or decision we are explaining, who are the target audience and do we want to optimize for trust, for understandability or do we wish to simply comply by legislation | Bussone *et al.* (2015), Dodge *et al.* (2019), Ehsan *et al.* (2019), Oh *et al.* (2018), Putnam and Conati (2019), Wang *et al.* (2019), Xie *et al.* (2019) |
| When to explain | 2. Provide explanations on demand, not all the time | For certain decisions and in certain moments users' may be interested in seeing more information on AI system decisions. However, constant display of full XAI documentation can hurt the user experience | Chazette and Schneider (2020), Cramer *et al.* (2008), Lim *et al.* (2009), Lim and Dey (2009), Oh *et al.* (2018) |
| How to explain | 3. Personalize explanations | There are various kinds of people with different levels of understanding of AI systems and XAI needs. This could be taken into account when explaining the system | Chazette and Schneider (2020), Cramer *et al.* (2008), Dodge *et al.* (2019), Weitz *et al.* (2019a), Wang *et al.* (2019), Xie *et al.* (2019) |
| | 4. Consider visualizing explanations | Users tend to anthropomorphize AI and may benefit from human-like explanations. Visualizing explanations may help some users to accept the AI system and its decisions better | Ngo *et al.* (2020), Schrills and Franke (2020), Weitz *et al.* (2019a), Weitz *et al.* (2019b) |
| | 5. Acknowledge the existence of trade-offs | For example, optimizing explanations for understandability can lead to less details, which can hurt end users' confidence in the explanation | Cheng *et al.* (2019), Dodge *et al.* (2019), Ehsan *et al.* (2019), Weitz *et al.* (2019a) |
| | 6. Consider potential misconceptions | Users may end up forming or having formed misconceptions regarding the AI system. These may shape behavior and interpretation of explanations in a certain way. Explanations that are able to reshape misconceptions in a constructive way of conceptual change are valuable | Cramer *et al.* (2008), Oh *et al.* (2018), Xie *et al.* (2019) |
| | 7. Link explanations to users' mental models | This makes the AI system easier to understand for end users, increasing transparency | Ngo *et al.* (2020), Lim *et al.* (2009) |
| | 8. Strengthen users' curiosity towards the system | To increase user satisfaction especially in creative and learning contexts, provide interesting and even surprising elements to keep the users' curiosity at a high level | Oh *et al.* (2018), Putnam and Conati (2019) |
| | 9. Ensure the visibility and discoverability of explanations | Make sure AI system end users find and become aware of explanations | Eslami *et al.* (2018) |

**Table 5.**
Recommendations for designing AI system explanations for end users

(*continued*)

| Recommendation categories | Design recommendation | Reasoning | Sources |
|---|---|---|---|
| | 10. Use metaphors to demystify how AI systems work | Metaphors can be more useful in increasing end users' understanding of AI systems than precise but difficult technical language | Ngo *et al.* (2020) |
| | 11. Support users' own thinking | In professional contexts, such as in medicine, the AI system should provide counterfactuals and explanations so users can reflect on and test their own thinking and hypotheses | Wang *et al.* (2019) |
| | 12. Provide access to source data | Especially in high-stakes decision making, such as in justice or in medicine, users may want to request access to raw data to build their trust in the AI system | Wang *et al.* (2019) |
| | 13. Provide users with generalized explanations rather than case-based explanations | Users may consider it quirky if the decision is explained to them with a particular event from the past. To increase user acceptance, refer to generalized past events instead | van der Waa *et al.* (2020) |
| What to explain | 14. Consider what part of the AI system to explain | Depending on the situation, users may wish to know more about, for example: (1) inputs; (2) outputs; (3) application; (4) situation; (5) model; (6) certainty; and (7) control | Broekens *et al.* (2010), Lim and Dey (2009) |
| | 15. Explain unfavorable decisions | Users are likely to demand explanations when they disagree with the system | Putnam and Conati (2019) |
| | 16. Communicate the uncertainties involved in the system's decision making | If there is a mismatch between users' expectations of the AI system and its actual capabilities, it hinders users' acceptance and trust building in it. Users should understand the risks of the AI system's making errors | Brennen (2020), Wang *et al.* (2019), Yin *et al.* (2019) |

**Table 5.**

Importantly, the identified recommendations are not universal, as evidenced by the first recommendation #1, as well as the findings from this literature review. There were multiple situations in which AI systems were explained to end users, and there were individual differences regarding end users' prior knowledge of AI systems and their ability to understand explanations (Dodge *et al.*, 2019; Xie *et al.*, 2019). As a solution, adding personalized explanations has been suggested (Dodge *et al.*, 2019; Wang *et al.*, 2019; Xie *et al.*, 2019). The reasoning for this included that end users vary in terms of their knowledge and understanding of AI systems and concerning when and what kind of explanations they need. Several studies (Chazette and Schneider, 2020; Cramer *et al.*, 2008; Lim *et al.*, 2009; Lim and Dey, 2009; Oh *et al.*, 2018) argued that always displaying explanations to end users would reduce the usability and even understandability of AI systems and that it would be counterproductive to force explanations to all AI systems. Thus, AI system designers should consider the UI design as a key component for understandability and test for most effective ways to display and visualize explanations (Eslami *et al.*, 2018; Ngo *et al.*, 2020; Schrills and Franke, 2020; Weitz *et al.*, 2019a, b).

When devising explanations, the research shows there are trade-offs with regards to what to focus on (Cheng *et al.*, 2019; Dodge *et al.*, 2019; Ehsan *et al.*, 2019; Weitz *et al.*, 2019a). Using metaphors to make the AI system more understandable (Ngo *et al.*, 2020) can backfire, as users may form inaccurate conceptions of how ML systems work over time (Cramer *et al.*, 2008; Oh *et al.*, 2018). Due to ubiquitous misconceptions about ML systems, one of the recommendations was to consider them when providing AI system explanations, with the goal of correcting the misconceptions (Cramer *et al.*, 2008; Oh *et al.*, 2018). One of the more creative uses of AI explanations for end users was given in the context of medicine, where the explanations may help practitioners increase their understanding of the underlying phenomena, adding value to decision making beyond what the model delivers (Wang *et al.*, 2019). However, Wang *et al.* (2019) also noted that giving users full access to the source data might increase dangers, such as extracting sensitive information from the source dataset or reverse engineering the ML model. Thus, explanations with such a high level of transparency are not suitable for all cases.

### *4.4 Future research agenda*

We extracted the future research directions from the reviewed studies as such, but also synthesized the literature to identify research directions more broadly in the field of XAI for end users. Most of the future research directions explicitly mentioned in our sample of studies were related to improving the empirical research setup of the work, such as (1) improvements specific to the research problem more generally augmenting the usability of the UI of the research tool (e.g. Broekens *et al.*, 2010; Chazette and Schneider, 2020) and (2) repeating the research setup elsewhere for increased reliability and proving reproducibility (e.g. Cirqueira *et al.*, 2020; Ngo *et al.*, 2020). The remaining explicitly stated future research directions can be divided into two main groups. The first group relates to research directions specific to the five objectives of XAI for end users discussed in Section 4.2. The second consists of research directions that are generally applicable across the five objectives of AI communication for end users.

The future research recommendations related to XAI for end users are presented in Table 6. Concerning end users' understanding of AI systems, future research should compare

**Table 6.**
Future research agenda concerning the goals and objectives of AI system explanations for end users

| Goal/objective | Future research direction | Source |
|---|---|---|
| Understandability | Investigate different groups of end users and their ability to understand AI system explanations | Cheng *et al.* (2019) |
| | Elucidate and determine the desired levels of understanding of AI systems of different stakeholders | This study |
| Trustworthiness | Investigate if involving humans in the loop of AI decision making can increase trust in AI systems | Cheng *et al.* (2019) |
| | Investigate emotional and cognitive factors involved in trust such as surprise, confusion and cognitive dissonance | Yin *et al.* (2019) |
| | Determine how various explanation types influence the resulting trust toward the AI system | This study |
| Transparency | Investigate the link between transparency and trust | Eiband *et al.* (2018) |
| | Investigate how information disclosure and presentation are linked to end users' perceived transparency of the explanations | This study |
| Controllability | Connect the goal of controllability to understandability, transparency, trustworthiness and fairness of the system | This study |
| Fairness | Approach the issue from various psychology of justice theories such as interactional justice | Binns *et al.* (2018) |
| | Studies regarding the fairness of AI system explanations could focus on how well end users understand the real behavior of the system based on provided explanations | This study |

various subgroups of end users and discern how individual differences influence an understanding of AI system explanations (Cheng *et al.*, 2019). In addition, there is a need to study at what level different stakeholders groups need to understand the AI systems they use. Regarding the trustworthiness of the system, there were a few suggestions. First would be to explore whether showcasing the presence of humans in the decision loop increased the perceived trustworthiness of the system (Cheng *et al.*, 2019). A second approach could be to investigate the various components of trust (i.e. emotional and cognitive) as well as situational aspects such as surprise and confusion (Yin *et al.*, 2019). Third, there is a need to study the different explanation types in further detail (e.g. rule-based vs example-based).

No explicit future research directions were given in the sample of studies. However, through synthesis of the literature we suggest future work to focus on further probing the connections between controllability and transparency, fairness, trustworthiness and understanding. This would help better frame controllability as a goal for AI system explanations for end users. Regarding the research on transparency, future research should explore the link between the transparency of the system and trust (Eiband *et al.*, 2018). Finally, for research on fairness, future research should focus on applying the psychology of justice theories (Binns *et al.*, 2018). In addition, research could investigate the link between understanding the system and end users' perceived fairness of it.

Table 7 summarizes the general future research directions presented by the studies. The research avenue that overwhelmingly most often appears validates the findings of the studies in real-world contexts. This speaks of the multiple experimental scenarios created to study XAI for end users and of the lack of real-world implementations of the proposed systems. Thus, the future research agenda of this domain must focus on field experiments, industry collaboration, and the study of real-world systems. Another general future research avenue that appeared more than once was to consider various XAI stakeholder groups (Binns *et al.*, 2018; Brennen, 2020). Previous work still seems to focus overwhelmingly on data scientists (Binns *et al.*, 2018; Brennen, 2020). However, research is also done on end users, as evident by this study and other stakeholder groups identified by Meske *et al.* (2022). Future research

| Future research direction | Sources |
|---|---|
| Validate the findings in real-world scenarios | Chazette and Schneider (2020), Cirqueira *et al.* (2020), Cheng *et al.* (2019), Eiband *et al.* (2018), Eslami *et al.* (2018), Lim and Dey (2009), Ngo *et al.* (2020) |
| Consider XAI aimed at various stakeholders | Binns *et al.* (2018), Brennen (2020) |
| Clarify XAI terminology and conceptualizations | Chazette and Schneider (2020) |
| Explore interactions between AI explanations and other design aspects | Chazette and Schneider (2020) |
| Investigate how end users' focus on explanations change over time, i.e. whether they at first focus more on whether they can trust the system and later other aspects | Cramer *et al.* (2008) |
| Explore the impacts of allowing users to question the AI system's decisions | Ehsan *et al.* (2019) |
| Investigate and elucidate industry- and profession-specific explanation needs | This study |
| Investigate the role of end users' education and understanding on understanding provided explanations | This study |
| Focus on explaining the dark side and unwanted consequences of AI systems | This study |

**Table 7.**
Future research agenda concerning the current general research profile of XAI and AI explanations for end users

agendas could even divide the end users into clusters based on, for example, the type of AI system used or individual differences.

Interestingly, Cramer *et al.* (2008) suggested investigating how end users' perceptions of AI systems change over time and whether they initially desire more information about whether they can trust the system and later something else. However, our literature review showed this suggestion remains unexplored. Another interesting future research avenue was to explore the outcomes of adding interactivity to the explanations, particularly in the form of enabling end users to question decisions (Ehsan *et al.*, 2019). Chazette and Schenider (2020) delivered two suggestions: to clarify the concepts and terminology involved in this research field and to explore the interactions between XAI for end users and other design aspects, such as the usability of the system. Interestingly, the education aspect was still largely missing from the body of literature, indicating that intervention studies looking into how education on AI systems influence end users' perceptions regarding provided explanations is needed. Finally, there is a growing body of IS research on the dark sides of AI-agents and AI systems (e.g. Cheng *et al.*, 2021) and explaining the negative and unwanted consequences of AI systems are largely absent in the literature synthesized in this study. Thus, we encourage future work to investigate how to explain unwanted consequences of AI systems to end users.

## 5. Discussion

### *5.1 Key findings*

We summarize our findings by answering our three research questions as follows.

*RQ1.* What are the goals and objectives of AI system explanations for end users?

We identified five high-level aims/goals for XAI and AI system explanations for end users. These were (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) fairness. These themes were interlinked. For example, higher transparency was found to support users' trust in the system (Ehsan *et al.*, 2019; Xie *et al.*, 2019; Wang *et al.*, 2019) and their overall understanding of it (Cramer *et al.*, 2008). In addition, the understandability of AI system explanations was associated with the fairness of the system itself (Binns *et al.*, 2018; Dodge *et al.*, 2019).

*RQ2.* What recommendations does the extant literature suggest for designing explanations for AI systems that facilitate positive outcomes?

We identified 16 unique design recommendations classified into four categories: general recommendations and recommendations addressing "when," "what" and "how" to explain AI system decisions. These are displayed in Table 5. The literature shows that there is no single way to explain AI systems for end users (Bussone *et al.*, 2015; Dodge *et al.*, 2019; Ehsan *et al.*, 2019; Oh *et al.*, 2018; Putnam and Conati, 2019; Wang *et al.*, 2019; Xie *et al.*, 2019). As the end users originate from various backgrounds and have different conceptions about ML models, the literature suggests personalized explanations (Chazette and Schneider, 2020; Cramer *et al.*, 2008; Dodge *et al.*, 2019; Weitz *et al.*, 2019a; Wang *et al.*, 2019; Xie *et al.*, 2019) and linking the explanations to users' existing conceptions and mental models (Ngo *et al.*, 2020; Lim *et al.*, 2009). Other often-mentioned design recommendations included offering explanations on demand (Chazette and Schneider, 2020; Cramer *et al.*, 2008; Lim *et al.*, 2009; Lim and Dey, 2009; Oh *et al.*, 2018), visualizing explanations (Schrills and Franke, 2020; Weitz *et al.*, 2019a, b), ensuring the visibility of the explanations in the UI (Eslami *et al.*, 2018), and clarifying the user system decision based on Ngo *et al.* (2020).

*RQ3.* What future research directions arise from the extant literature?

Our analysis of the future research areas suggested by the reviewed studies revealed three main future research directions. First, there is a need to validate the findings of the studies in

real-world contexts (Chazette and Schneider, 2020; Cirqueira *et al.*, 2020; Cheng *et al.*, 2019; Eiband *et al.*, 2018; Eslami *et al.*, 2018; Lim and Dey, 2009; Ngo *et al.*, 2020). This includes field experiments and research with real-world systems. Second, various studies have discovered individual differences between AI end users, and the differences need to be better understood (Ngo *et al.*, 2020; Lim *et al.*, 2009). These differences link to creating personalized AI system explanations, as well as educating the masses (Chazette and Schneider, 2020; Cramer *et al.*, 2008; Dodge *et al.*, 2019). Third and finally, in our systematic review, we observed that the participant samples were not adequately described in several studies (See Table A1). This is a shortcoming in the XAI for end users' research field. As the end users' perceptions played a primary role in various studies, and there was evidence of individual differences (Ngo *et al.*, 2020; Lim *et al.*, 2009), understanding who exactly the end users are should be paramount. Thus, future research should particularly emphasize the rigorous sampling of research subjects and the detailed reporting of the profile and characteristics of the samples.

### *5.2 Research implications*

The current study makes three principal contributions to XAI research. First, we respond to the recent calls for (1) research on the XAI area from an HCI perspective (Brennen, 2020) and (2) increased focus on AI system end users (Weitz *et al.*, 2019b) by systematically reviewing and synthesizing the existing literature. Our work also supports literature reviews carried out on the technical aspect of XAI (Antoniadi *et al.*, 2021; Arrieta *et al.*, 2020) by extending the current body of knowledge on the explainability needs and goals of end users. Serving these needs will ultimately be the task of system developers. This review can help XAI developers and system designers in eliciting design requirements.

Second, through our review of the literature, we have elucidated findings of five objectives of XAI for end users, namely (1) understandability, (2) trustworthiness, (3) transparency, (4) controllability and (5) fairness. In doing so, the current study extends on Meske *et al.* (2022), who presented five objectives of XAI for AI system developers. Importantly, the five objectives identified here have connections to those of developers. For example, while developers must understand AI systems to develop and operate them (Anjomshoae *et al.*, 2019; Antoniadi *et al.*, 2021), it is important for end users to feel in control of the system (Ngo *et al.*, 2020; Oh *et al.*, 2018; Wang *et al.*, 2019). Hence, the developers must create tools and systems that enable end users to understand the system they are using and feel control of it. Moreover, this connects to the discussion on the role of XAI as an element of AI governance (Minkkinen *et al.*, 2020b, 2022a; Mäntymäki *et al.*, 2022; Seppälä *et al.*, 2021). According to Seppälä *et al.* (2021), AI design and development play a significant role in translating ethical principles into governed AI, while explainability in turn, is one of the key design issues related to AI governance.

Third, by synthesizing the future research areas suggested by the reviewed studies, we have provided a future research agenda for the research of XAI and AI communication for end users. Regarding future research directions, one of the primary pursuits should be to validate the findings of the studies in real-world scenarios (Chazette and Schneider, 2020; Cirqueira *et al.*, 2020; Cheng *et al.*, 2019; Eiband *et al.*, 2018; Eslami *et al.*, 2018; Lim and Dey, 2009; Ngo *et al.*, 2020). While AI system development seems to advance primarily through the efforts of the industry, academia seems to be at the forefront concerning AI communication for end users. This is evident in the promising results of the research reviewed in this study.

### *5.3 Practical implications*

The key practical implications and contributions of this review are the summary of the design recommendations presented in Table 5. The presented list of 16 design recommendations is a summary of the reviewed 25 studies and may be used by practitioners and designers as a

checklist of what to consider when presenting AI systems' explanations to end users. Based on the results displayed in Table 5, we developed a framework for designing XAI and AI system communication for end users (Figure 3). The two ubiquitously applicable steps of researching the context and providing explanations on demand are followed by a set of "how to explain" and "what to explain" considerations, which, while not universally relevant to all AI systems, can be useful for AI communication designers.

Several studies have stated there is a need to implement their findings into practice (Chazette and Schneider, 2020; Cirqueira *et al.*, 2020; Cheng *et al.*, 2019; Eiband *et al.*, 2018; Eslami *et al.*, 2018; Lim and Dey, 2009; Ngo *et al.*, 2020). This demonstrates how while academic research has produced several ideas for various scenarios on how to improve XAI for end users, real world scenario validation is still lacking. Hence, the next step is for practitioners to adopt them into use. While the research findings seem promising concerning being able to, for example, increase users' trust in AI systems (Ehsan *et al.*, 2019; Schrills and Franke, 2020; Wang *et al.*, 2019; Weitz *et al.*, 2019b; Xie *et al.*, 2019; Yin *et al.*, 2019) and feeling of control (Ngo *et al.*, 2020; Oh *et al.*, 2018: Wang *et al.*, 2019) of the AI system, these results can be different under real-world circumstances. Thus, practitioners are encouraged to take promising design recommendations and adapt them into practice, but measure their effects on end users. For example, it is worthy to investigate whether AI system communication has the potential to alleviate trust issues that end users face with AI systems (Zarifis *et al.*, 2020) or technostress (Tarafdar *et al.*, 2020). Continuous measurement and feedback are important, as each system and use case are unique. Furthermore, blindly trusting findings from any usability research in the XAI field would be counterproductive due to the novelty and formative state of the research area.

The lack of explainability of ML models and resulting challenges with governance have been identified as one of the biggest obstacles to adopting them into use (e.g. Adadi and Berrada, 2018; Došilović *et al.*, 2018; Rana *et al.*, 2021; Weitz *et al.*, 2019a). With our work, we present key design recommendations regarding AI system end users' explanation needs. This has implications for XAI developers who seek to create visualizations and other forms of XAI to support system users, as well as for system and UI engineers who aim to present explanations in a way that end users find useful. The findings of this study pave the way for creating more transparent and clear explanations of AI systems, which can also negate some of the unintended consequences and inhibitors for adopting AI systems that companies as well as individuals face (Rana *et al.*, 2021).

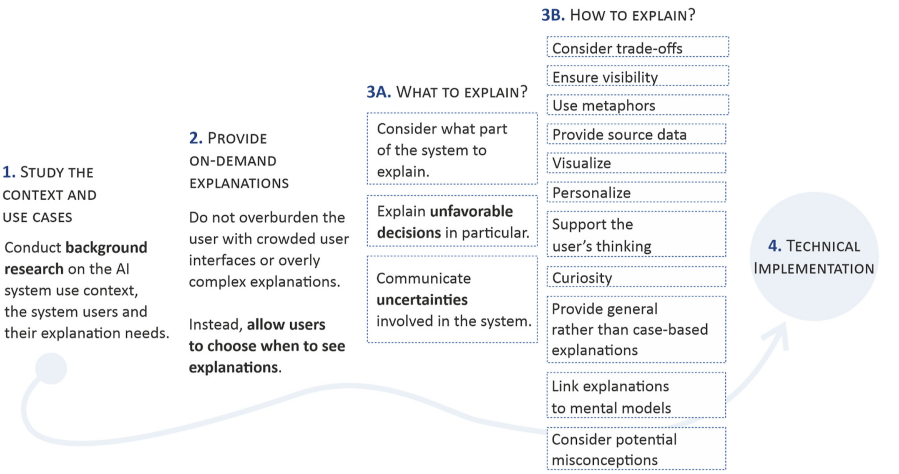


**Figure 3.**
A design framework for AI system communication for end users that fits the five goals of understandability, trustworthiness, transparency, controllability and fairness

### 5.4 Limitations

The limitations of the current study relate to two main areas: (1) the literature search and (2) data extraction and analysis. In the literature search process, we focused on two widely used bibliometric databases: Web of Science and Scopus. However, there are still some venues and publications that are not indexed in these databases. Hence, despite the backward chaining (Wohlin, 2014) executed, there is a risk of missing some studies.

In addition, we focused exclusively on peer-reviewed studies and omitted gray literature. This was done to ensure the quality of the material included in the analysis but again has the drawback of potentially missing some relevant information. The initial literature search process was conducted by a single author. Accordingly, there is a minor risk of false negatives in the sample. False positives should not exist, since all authors participated in reviewing the final sample.

With regards to the data extraction and analysis, we extracted data to focus on answering our two research questions. However, unlike some empirical quantitative studies with similar research setups, the studies included in the final review were heterogeneous in terms of contextual coverage and thus, to some extent, challenging to compare. For example, the contextual coverage of the reviewed studies ranged from medical AI systems (Xie *et al.*, 2019) to online advertising (Eslami *et al.*, 2018). While the contexts and research setups were different (see Table A1 for more information), the core issues the studies dealt with were the same. Nonetheless, the contextual and methodological variance between the reviewed studies must be acknowledged as a potential limitation.

## 6. Conclusion

We conducted a systematic literature review on AI system explanations for end users. The systematic literature search and selection process resulted in 25 empirical research articles, which were looked into in detail in this study. Through our main findings, this work makes three key contributions. First, we identified and elucidated the objectives and goals of AI communication for end users. Second, we extracted, analyzed and further developed design recommendations for explaining AI systems to end users. Third, based on the findings, we produced a synthesized design framework for end-user AI communication (Figure 3) which serves as a structured form of the discovered design recommendations. The framework provides AI system communication designers and XAI specialists a solid starting point for understanding the explanation needs of end users, and provides suggestions on how to communicate about the AI system in a way that is understandable, trustworthy, transparent, controllable and fair.

## Notes

1. The founder of Twitter, Jack Dorsey, has repeatedly communicated how the (partially automated) moderation practices of Twitter should be made more transparent, available at: https://economictimes.indiatimes.com/tech/tech-bytes/twitter-intends-to-make-its-content-moderation-practices-more-transparent-jack-dorsey/articleshow/81223668.cms (November 20, 2021).
2. SHapley Additive exPlanations (SHAP), available at: https://shap.readthedocs.io/en/latest/index.html (accessed April 2, 2022).

## References

Adadi, A. and Berrada, M. (2018), "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)", *IEEE Access*, Vol. 6, pp. 52138-52160, doi: 10.1109/ACCESS.2018.2870052.

Anjomshoae, S., Najjar, A., Calvaresi, D. and Främling, K. (2019), "Explainable agents and robots: results from a systematic literature review", in *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1078-1088.

Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A. and Mooney, C. (2021), "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review", *Applied Sciences*, Vol. 11 No. 11, doi: 10.3390/app11115088.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Lopez, S., Molina, D., Benjamin, R., Chatila, R. and Herrera, F. (2020), "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities, and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115, doi: 10.1016/j.inffus.2019.12.012.

Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. (2020), "Challenges of explaining the behavior of blackbox AI systems", *MIS Quarterly Executive*, Vol. 19 No. 4, pp. 259-278, doi: 10.17705/2msqe.00037.

Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T. and Salovaara, A. (2021), "Sociotechnical envelopment of artificial intelligence: an approach to organizational deployment of inscrutable artificial intelligence systems", *Journal of the Association for Information Systems*, Vol. 22 No. 2, doi: 10.17705/1jais.00664.

Batmaz, Z., Yurekli, A., Bilge, A. and Kaleli, C. (2019), "A review on deep learning for recommender systems: challenges and remedies", *Artificial Intelligence Review*, Vol. 52 No. 1, pp. 1-37, doi: 10.1007/s10462-018-9654-y.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J. and 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*, pp. 1-14, doi: 10.1145/3173574.3173951.

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*, pp. 1-7, doi: 10.1145/3334480.3383047.

Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C. and Meyer, J.J. (2010), "Do you get it? User-evaluated explainable BDI agents", in *Proceedings of the German Conference on Multiagent System Technologies*, pp. 28-39, doi: 10.1007/978-3-642-16178-0_5.

Burrell, J. (2016), "How the machine "thinks": understanding opacity in machine learning algorithms", *Big Data and Society*, Vol. 3 No. 1, doi: 10.1177/2053951715622512.

Bussone, A., Stumpf, S. and O'Sullivan, D. (2015), "The role of explanations on trust and reliance in clinical decision support systems", in *Proceedings of the 2015 International Conference on Healthcare Informatics*, pp. 160-169, doi: 10.1109/ICHI.2015.26.

Cai, C.J., Winter, S., Steiner, D., Wilcox, L. and Terry, M. (2019), "Hello AI": uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making", in *Proceedings of the ACM on Human–Computer Interaction*, Vol. 3, pp. 1-24, doi: 10.1145/3359206.

Chazette, L. and Schneider, K. (2020), "Explainability as a non-functional requirement: challenges and recommendations", *Requirements Engineering*, Vol. 25 No. 4, pp. 493-514, doi: 10.1007/s00766-020-00333-1.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M. and 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*, pp. 1-12, doi: 10.1145/3290605.3300789.

Cirqueira, D., Nedbal, D., Helfert, M. and Bezbradica, M. (2020), "Scenario-based requirements elicitation for user-centric explainable AI", in *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 321-341, doi: 10.1007/978-3-030-57321-8_18.

Collins, C., Dennehy, D., Conboy, K. and Mikalef, P. (2021), "Artificial intelligence in information systems research: a systematic literature review and research agenda", *International Journal of Information Management*, Vol. 60, doi: 10.1016/j.ijinfomgt.2021.102383.

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L. and 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 No. 5, doi: 10.1007/s11257-008-9051-3.

Dawes, S. (2021), "An autonomous robot may have already killed people – here's how the weapons could be more destabilizing than nukes", available at: https://theconversation.com/an-autonomous-robot-may-have-already-killed-people-heres-how-the-weapons-could-be-more-destabilizing-than-nukes-168049 (accessed 2 April 2022).

Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K. and Dugan, C. (2019), "Explaining models: an empirical study of how explanations impact fairness judgment", in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 275-285, doi: 10.1145/3301275.3302310.

Dogan, O., Tiwari, S., Jabbar, M.A. and Guggari, S. (2021), "A systematic review on AI/ML approaches against COVID-19 outbreak", *Complex and Intelligent Systems*, pp. 1-24, doi: 10.1007/s40747-021-00424-8.

Došilović, F.K., Brčić, M. and Hlupić, N. (2018), "Explainable artificial intelligence: a survey", in *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics, and Microelectronics (MIPRO)*, pp. 210-215, doi: 10.23919/MIPRO.2018.8400040.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B. and Riedl, M.O. (2019), "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*, pp. 263-274, doi: 10.1145/3301275.3302316.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M. and Hussmann, H. (2018), "Bringing transparency design into practice", in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pp. 211-223, doi: 10.1145/3172944.3172961.

Eslami, M., Krishna Kumaran, S.R., Sandvig, C. and Karahalios, K. (2018), "Communicating algorithmic process in online behavioral advertising", in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-13, doi: 10.1145/3173574.3174006.

European Commission (2020), "White paper on artificial intelligence—a European approach to excellence and trust", available at: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed 28 March 2022).

Evans, R., Bošnjak, M., Buesing, L., Ellis, K., Pfau, D., Kohli, P. and Sergot, M. (2021), "Making sense of raw input", *Artificial Intelligence*, Vol. 299, doi: 10.1016/j.artint.2021.103521.

Hornung, O. and Smolnik, S. (2021), "AI invading the workplace: negative emotions towards the organizational use of personal virtual assistants", *Electronic Markets*, pp. 1-16, doi: 10.1007/s12525-021-00493-0.

Jobin, A., Lenca, M. and Vayena, E. (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, Vol. 1 No. 9, pp. 389-399, doi: 10.1038/s42256-019-0088-2.

Karamitsos, I., Albarhami, S. and Apostolopoulos, C. (2020), "Applying DevOps practices of continuous automation for machine learning", *Information*, Vol. 11 No. 7, doi: 10.3390/info11070363.

Kitchenham, B. and Brereton, P. (2013), "A systematic review of systematic review process research in software engineering", *Information and Software Technology*, Vol. 55 No. 12, pp. 2049-2075, doi: 10.1016/j.infsof.2013.07.010.

Kuwajima, H., Tanaka, M. and Okutomi, M. (2019), "Improving transparency of deep neural inference process", *Progress in Artificial Intelligence*, Vol. 8 No. 2, pp. 273-285, doi: 10.1007/s13748-019-00179-x.

Laato, S., Mäntymäki, M., Birkstedt, T., Islam, A.K.M. and Hyrynsalmi, S. (2021), "Digital transformation of software development: implications for the future of work", in *Proceedings of the Conference on e-Business, e-Services and e-Society*, pp. 609-621, doi: 10.1007/978-3-030-85447-8_50.

Laato, S., Birkstedt, T., Islam, A.K.M., Hyrynsalmi, S. and Mäntymäki, M. (2022), "Trends and trajectories in the software industry: implications for the future of work", *Information Systems Frontiers*. doi: 10.1007/s10796-022-10267-4 (In press).

Liang, T.P., Robert, L., Sarker, S., Cheung, C.M., Matt, C., Trenz, M. and Turel, O. (2021), "Artificial intelligence and robots in individuals' lives: how to align technological possibilities and ethical issues", *Internet Research*, Vol. 31 No. 1, pp. 1-10, doi: 10.1108/INTR-11-2020-0668.

Lim, B. and Dey, A. (2009), "Assessing demand for intelligibility in context-aware applications", in *Proceedings of the 11th International Conference on Ubiquitous Computing*, pp. 195-204, doi: 10.1145/1620545.1620576.

Lim, B., Dey, A. and 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*, pp. 2119-2128, doi: 10.1145/1518701.1519023.

Mäntymäki, M., Minkkinen, M., Birkstedt, T. and Viljanen, M. (2022), "Defining organizational AI governance", *AI and Ethics*, pp. 1-7, doi: 10.1007/s43681-022-00143-x.

Meske, C., Bunde, E., Schneider, J. and Gersch, M. (2022), "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities", *Information Systems Management*, Vol. 39 No. 1, pp. 53-63, doi: 10.1080/10580530.2020.1849465.

Ming, Y., Xu, P., Cheng, F., Qu, H. and Ren, L. (2019), "ProtoSteer: steering deep sequence model with prototypes", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 26 No. 1, pp. 238-248, doi: 10.1109/TVCG.2019.2934267.

Minkkinen, M., Niukkanen, A. and Mäntymäki, M. (2022a), "What about investors? ESG analyses as tools for ethics-based AI auditing", *AI and Society*. doi: 10.1007/s00146-022-01415-0.

Minkkinen, M., Zimmer, M.P. and Mäntymäki, M. (2022b), "Co-shaping an ecosystem for responsible AI: an analysis of expectation work in response to a technological frame", *Information Systems Frontiers*. doi: 10.1007/s10796-022-10269-2 (in press).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019), "Model cards for model reporting", in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229, doi: 10.1145/3287560.3287596.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and Prisma Group (2009), "Preferred reporting items for systematic reviews and meta-analyzes: the PRISMA statement", *PLoS Medicine*, Vol. 6 No. 7, doi: 10.7326/0003-4819-151-4-200908180-00135.

Morschheuser, B., Hamari, J., Koivisto, J. and Maedche, A. (2017), "Gamified crowdsourcing: conceptualization, literature review, and future agenda", *International Journal of Human-Computer Studies*, Vol. 106, pp. 26-43, doi: 10.1016/j.ijhcs.2017.04.005.

Ngo, T., Kunkel, J. and 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*, pp. 183-191, doi: 10.1145/3340631.3394841.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S. and Suh, B. (2018), "I lead, you help, but only with enough details: understanding user experience of co-creation with artificial intelligence", in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1-13, doi: 10.1145/3173574.3174223.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A.K. (2020), "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis", *Accident Analysis and Prevention*, Vol. 136, doi: 10.1016/j.aap.2019.105405.

Putnam, V. and Conati, C. (2019), "Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS)", in *Proceedings of IUI Workshops*, Vol. 19.

Rana, N.P., Chatterjee, S., Dwivedi, Y.K. and Akter, S. (2021), "Understanding dark side of artificial intelligence (AI) integrated business analytics: assessing firm's operational inefficiency and competitiveness", *European Journal of Information Systems*, pp. 1-24, doi: 10.1080/0960085X.2021.1955628.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016), ""Why should I trust you?" Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, doi: 10.1145/2939672.2939778.

Schrills, T. and Franke, T. (2020), "Color for characters-effects of visual explanations of AI on trust and observability", in *Proceedings of the International Conference on Human-Computer Interaction*, pp. 121-135, doi: 10.1007/978-3-030-50334-5_8.

Seppälä, A., Birkstedt, T. and Mäntymäki, M. (2021), "From ethical principles to governed AI", in *Proceedings of the 42nd International Conference on Information Systems (ICIS2021)*.

Shneiderman, B. (2020), "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems", *ACM Transactions on Interactive Intelligent Systems*, Vol. 10 No. 4, pp. 1-31, doi: 10.1145/3419764.

Smith, G. (2021), "An epic failure: overstated AI claims in medicine - independent investigations are finding that AI algorithms used in hospitals are not all they claim to be", available at: https://mindmatters.ai/2021/08/an-epic-failure-overstated-ai-claims-in-medicine (accessed 2 April 2022).

Strauss, A. and Corbin, J. (1998), *Basics of Qualitative Research Techniques*, Sage Publications, Thousand Oaks, CA.

Tarafdar, M., Beath, C.M. and Ross, J.W. (2019), "Using AI to enhance business operations", *MIT Sloan Management Review*, Vol. 60 No. 4, pp. 37-44.

Tarafdar, M., Maier, C., Laumer, S. and Weitzel, T. (2020), "Explaining the link between technostress and technology addiction for social networking sites: a study of distraction as a coping behavior", *Information Systems Journal*, Vol. 30 No. 1, pp. 96-124, doi: 10.1111/isj.12253.

van der Waa, J., Schoonderwoerd, T., van Diggelen, J. and Neerincx, M. (2020), "Interpretable confidence measures for decision support systems", *International Journal of Human–Computer Studies*, Vol. 144, doi: 10.1016/j.ijhcs.2020.102493.

van der Waa, J., Nieuwburg, E., Cremers, A. and Neerincx, M. (2021), "Evaluating XAI: a comparison of rule-based and example-based explanations", *Artificial Intelligence*, Vol. 291, doi: 10.1016/j.artint.2020.103404.

von Eschenbach, W.J. (2021), "Transparency and the black box problem: why we do not trust AI", *Philosophy and Technology*, Vol. 34, pp. 1607-1622, doi: 10.1007/s13347-021-00477-0.

Wang, D., Yang, Q., Abdul, A. and Lim, B.Y. (2019), "Designing theory-driven user-centric explainable AI", in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-15, doi: 10.1145/3290605.3300831.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T. and André, E. (2019a), ""Let me explain!": exploring the potential of virtual agents in explainable AI interaction design", *Journal on Multimodal User Interfaces*, Vol. 15, pp. 87-98, doi: 10.1007/s12193-020-00332-0.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T. and André, E. (2019b), "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*, pp. 7-9, doi: 10.1145/3308532.3329441.

Weitz, K., Schlagowski, R. and André, E. (2021), "Demystifying artificial intelligence for end-users: findings from a participatory machine learning show", in *Proceedings of the German Conference on Artificial Intelligence (Künstliche Intelligenz)*, Vol. 12873, pp. 257-270, doi: 10.1007/978-3-030-87626-5_19.

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*, pp. 1-10, doi: 10.1145/2601248.2601268.

Xie, Y., Gao, G. and Chen, X.A. (2019), "Outlining the design space of explainable intelligent systems for medical diagnosis", in *Joint Proceedings of the ACM IUI 2019 Workshops*, pp. 1-8, doi: 10.48550/arXiv.1902.06019.

Yin, M., Wortman Vaughan, J. and Wallach, H. (2019), "Understanding the effect of accuracy on trust in machine learning models", in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-12, doi: 10.1145/3290605.3300509.

Zarifis, A., Kawalek, P. and Azadegan, A. (2020), "Evaluating if trust and personal information privacy concerns are barriers to using health insurance that explicitly utilizes AI", *Journal of Internet Commerce*, Vol. 20 No. 1, pp. 66-83, doi: 10.1080/15332861.2020.1832817.

Zhu, J., Liapis, A., Risi, S., Bidarra, R. and 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*. doi: 10.1109/CIG.2018.8490433.

## Appendix 1

### Search strings for Scopus and web of science

The search string for Scopus:
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 (DOC-TYPE, "le") OR LIMIT-TO (DOCTYPE, "no")) AND (LIMIT-TO (LAN-GAUGE, "English"))

The search string for the Web of Science Core Collection:
(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)

## Appendix 2

| Study | Study approach | Methodology | Data |
|---|---|---|---|
| Binns *et al.* (2018) | An experimentation into people's perceptions of justice in algorithmic decision making under different scenarios and explanation styles | Three studies: (1) A scenario-based in-lab user study. Data collected via expert interviews (2) A between-subjects study of five scenarios, with 65 participants in each scenario (3) A within-subjects study with 65 participants in loan and insurance cases | In-lab study: 19 participants from a small town in the UK The two online studies: 390 British participants over 18 years old recruited via Prolific Academic |
| Brennen (2020) | An exploration into what stakeholders want from XAI | Expert interviews | Company founders, investors, potential end users, and academia members ($n$ = 40). Presumably from the USA due to the authors' universities |
| Broekens *et al.* (2010) | An examination of the usefulness and naturalness of types of explanations for the actions of agents based on belief desire intention (BDI) | A between-subjects study of three conditions with 10 participants in each scenario | Study subjects were a "balanced mix of family, friends, colleagues, and students of the first two authors," with an average education level between bachelor's and master's. ($n$ = 30). Presumably from the Netherlands due to the authors' university |
| Bussone *et al.* (2015) | An exploration into how explanations are related to domain experts' trust and reliance on clinical decision support systems | An exploratory between-group user study employing two different versions of a CDSS prototype (comprehensive vs selective version). Using the think-aloud method, participants' decision making and trust when exploring the prototype was analyzed | Eight participants (seven primary care practitioners and one nurse), recruited through ads in medical network groups and forums, medical schools, and local primary care offices in the UK |
| Cai *et al.* (2019) | An investigation into what are the key types of information that medical experts want and need when introduced to a diagnostic AI assistant | A three-phase lab study with pathologists. Participants were interviewed before, during, and after presenting DNN predictions for prostate cancer diagnosis to explore the types of information they needed from the AI assistant | 21 pathologists participated in the study, recruited from a pool of remote contractors assisting Google Health with pathology projects. Presumably from the USA due to the authors' university |
| Chazette and Schneider (2020) | An exploration into what users see as the advantages and disadvantages of embedded explanations in software systems and what is their current level of transparency | An online questionnaire using LimeSurvey with 16 questions (11 multiple choice, five open-ended) on software skills and use, explanation needs and frequency, and the presentation of explanations | Snowball sampling with the help of the personal networks of the authors. Target population included adult end users of all ages, with different occupations. In total, 107 respondents completed the survey, of which 84% were from Brazil and 16% from Germany |
| Cheng *et al.* (2019) | An investigation of what kind of design principles would help non-expert stakeholders to understand how decision-making algorithms work | An online between-subjects study of five conditions, with each participant randomly assigned to a scenario, completed via Amazon MTurk | Participants of the study were recruited from Amazon MTurk. To qualify for the study participants needed to reside in the US, be aged 18 or above, and have a HIT approval rate of 90% or above. ($n$ = 199) |

(*continued*)

**Table A1.** Studies included in the analysis, their research approaches, methods and data

| Study | Study approach | Methodology | Data |
|---|---|---|---|
| Cirqueira *et al.* (2020) | A demonstration of the usage of scenario-based requirement elicitation for XAI in a fraud detection context | A problem-centered expert interview study to validate two fraud detection scenarios that could be adopted to identify expert requirements for adequate explanations | Three banking fraud specialists from one bank in Austria participated in the study, but the recruitment process was not provided |
| Cramer *et al.* (2008) | Examines the influence of transparency on users' trust and acceptance of content-based art recommendation systems | A between-subjects user study of three conditions with 22 participants in the first condition and 19 in the second and third | Participants were volunteers from the researchers' personal and professional networks, were relatively well educated, and had a good knowledge using computers. The participants' country of origin was not stated, presumably the Netherlands based on the authors' university |
| Dodge *et al.* (2019) | An exploration into how four types of programmatically created explanations affect people's fairness judgment of ML systems | An online survey with four explanation styles; each participant presented with six fairness judgment cases | 160 people from Amazon MTurk, with criteria that the participant must live in the US and have completed at least 1,000 tasks in MTurk with at least a 98% approval rate |
| Ehsan *et al.* (2019) | An investigation of how to train a neural rationale generator to create rationale styles and how people perceive them | Both between- and within-subjects user study with participants split into two equal groups with two identical experimental conditions, differing only by type of candidate rationale. The first group evaluated a focused-view rationale, whereas the second group had complete-view rationales. Participants were asked to view five videos with a set of rationales each and to rate each rationale based on four different statements | 128 participants were recruited through TurkPrime: 93% of the participants lived in the US, while the 7% that were left were from India |
| Eiband *et al.* (2018) | An explorative quest to advance existing UI guidelines for increased transparency and to improve users' mental models, with the particular case of Freeletics Bodyweight Application | A stage-based participatory process consisting of different phases: (1) semi-structured interviews of app users about their current mental models, (2) card sorting to find which components of the app users thought relevant for the perceived transparency of the app, (3) user testing of the prototype versions of the new UI, (4) evaluation of the prototypes with users | 14 active users of the app were recruited for the interviews in a park (presumably in Germany), and 11 users for the card sorting, a mixture of long- and short-term users of the app. The number of participants in the prototype testing was not stated. In the evaluation stage, 15 users participated. The participants in all stages were presumably from Germany, based on the authors' university |
| Eslami *et al.* (2018) | An investigation of how revealing users' parts of the algorithmic process affects their perceptions of online advertising and its platforms | A lab study in which users viewed the actual personalized ads and explanations the advertisers had given to them, followed by what an advertising algorithm had inferred about them. In the last phase, users created themselves advertising in an ad creation interface and wrote their own desired explanations for an ad of a product of interest | 32 participants from San Francisco, United States, and the surrounding area. Participants were picked from a larger group of interested people by non-probability modified quota sampling to balance five characteristics with the proportions of the US's population: gender, age, education, race/ethnicity and socioeconomic status |

**Table A1.** (*continued*)

| Study | Study approach | Methodology | Data |
|---|---|---|---|
| Lim and Dey (2009) | An experimentation into what kind of information demands and under which circumstances users have them using four real-world applications | Two experiments: (1) A between-subjects survey study in which participants were shown three to four scenarios of one application followed by two to five instances of the scenarios. Participants were asked to describe their feelings about the application and what kind of information needs they would have (2) A between-subjects study for intelligibility types, which was formulated based on the results of experiment one. Participants were assigned to a survey with only one intelligibility type. Participants were asked to rate their satisfaction with the application using a seven-point Likert scale and questions about the usefulness of explanations | (1) 250 participants in the first experiment recruited from Amazon Mechanical Turk. (2) 610 participants in the second experiment, recruited also from MTurk. Participants were distributed evenly across the 12 conditions The geographical distribution of the participants was not provided |
| Lim *et al.* (2009) | An examination into what kind of explanation types are the most effective to describe the workings of context-aware intelligent systems | Two experiments. In both, participants were allowed to explore the system, after which their understanding of the system was tested. (1) A between-subjects study with three conditions to explore the effectiveness of question types. (2) Same procedure as in the first, but combined with two additional conditions ("What If" and "How To") | (1) 53 participants in the first experiment, divided between the three conditions, and (2) 158 participants in the second experiment, divided (not evenly) among the five conditions. Recruitment procedure and country of the participants were not stated |
| Ngo *et al.* (2020) | An examination of users' mental models in using recommender systems, namely, Netflix | A semi-structured interview study focusing on participants' experience with Netflix. Participants were asked about the workings and data processing of Netflix and asked to draw their own image of Netflix | 10 interviewees with advanced experience with Netflix. Recruitment procedure and country of origin not stated |
| Oh *et al.* (2018) | An investigation to understand the user experience in art co-creation with AI | A between-subjects study with four conditions and a treatment condition. Participants performed a series of drawing tasks with a think-aloud method and were interviewed afterwards about their experience with the tool. Users' experience was also quantitatively measured with a survey | 30 participants were recruited through an announcement in Seoul National University's online community website (thus, they are presumed to be South Korean) |
| Putnam and Conati (2019) | A quest to understand whether and when it is necessary for an intelligent tutoring system to explain its underlying user modeling techniques to students | An user experiment in which participants studied the materials provided, did a pre-test based on the context of materials, and used an adaptive constraint satisfaction problem (CSP) applet to solve two CSPs, followed by a post-test questionnaire and interview | Nine participants (university students) recruited from an introductory AI course at a university in North America |

(*continued*)

**Table A1.**

| Study | Study approach | Methodology | Data |
|---|---|---|---|
| Schrills and Franke (2020) | An investigation of how prototypical visualization approaches aimed at increasing the explainability of ML systems affect users' perceived trustworthiness and observability of the system | An online within-subjects study with three conditions that presented different information visualizations. Users' agreement with the classification was measured after each stimulus | 83 participants were recruited via e-mail, social networks, and at the local university. Geographic distribution of the participants was not provided |
| van der Waa *et al.* (2020) | An investigation of what properties make a confidence measure desirable and why, and how an interpretable confidence measure (ICM) is interpreted by users | Two user experiments: (1) An interview study with domain experts to evaluate "transparency of the case-based reasoning approach underlying an ICM compared to other confidence measures" (2) Quantitative online survey with users to evaluate users' interests and preferences regarding the explanations provided by a decision-support system (autonomous car) regarding its confidence in its advice | (1) "Several domain experts" participated in the study. Recruitment process and country of origin not stated (2) 40 participants recruited via Amazon Mechanical Turk |
| Wang *et al.* (2019) | A quest for designing a conceptual framework for building human-centered, decision-theory-driven XAI, based on which an explainable clinical diagnostic tool for intensive care phenotyping was designed in co-creation with clinicians | A co-design lab study with clinicians. Participants were asked to use the diagnostics dashboard and diagnose patient cases using it. Sessions were recorded, and participants were instructed to think aloud during their diagnostic process | 14 medical professionals recruited from a local hospital. Country and background were not stated |
| Weitz *et al.* (2019a) | An exploration of how incorporating virtual agents into XAI designs affects the perceived trust of users | A between-subjects user study. Participants interacted with a graphical user interface and were split into four test groups with different types of visualizations and audio explanations. Participants were asked to rate their impressions and trust in the system | 60 participants. Recruitment process and background of the participants were not provided |
| Weitz *et al.* (2019b) | An examination of how using virtual agents in explanations affects the trustworthiness of autonomous intelligent systems | A between-subjects user study with two conditions. The first group received explanations from a virtual agent while the other received only visual explanations. Users' perceived trust of the system was measured afterwards with a questionnaire | 30 participants. Recruitment process and the participants' country of origin were not stated |
| Xie *et al.* (2019) | An exploration into what medical professionals consider as explainable when interacting with data for diagnosis and treatment purposes | An interview study consisting of questions revolving around the professionals' working practices, challenges, and experience using computer-based systems to facilitate medical work | Sample consisted of six medical professionals from California, US, recruited via online participant call |

**Table A1.** (*continued*)

| Study | Study approach | Methodology | Data |
|---|---|---|---|
| Yin *et al.* (2019) | An examination of whether people's trust in a model varies depending on the model's stated accuracy on "held-out data" and on its observed accuracy in practice | Three experiments: (1) A between-subjects study with 10 treatments. Users were randomly assigned to one of five accuracy levels and asked to make predictions about the outcomes of 40 speed dating events<br>(2) A between-subjects study with two sub-experiments and two conditions with different levels of observed accuracy. Users were again asked to predict the outcome of 40 speed dates<br>(3) A between-subjects study with six conditions varying along the stated accuracy and observed accuracy, again with the same 40 prediction tasks | There were 1,994 participants in the first experiment, 757 participants in the second, and 1,042 participants in the third. All participants were from the United States and recruited via Amazon MTurk |

**Table A1.**

**Corresponding author**
Samuli Laato can be contacted at: sadala@utu.fi