The Role of Explainable AI in the Research Field of AI Ethics

HEIDI VAINIO-PEKKA, MAMIA ORI-OTSE AGBESE, and MARIANNA JANTUNEN,
University of Jyväskylä, Finland
VILLE VAKKURI, University of Vaasa, Finland
TOMMI MIKKONEN, University of Jyväskylä, Finland
REBEKAH ROUSI, University of Vaasa, Finland
PEKKA ABRAHAMSSON, Tampere University, Finland

Ethics of Artificial Intelligence (AI) is a growing research field that has emerged in response to the challenges related to AI. Transparency poses a key challenge for implementing AI ethics in practice. One solution to transparency issues is AI systems that can explain their decisions. Explainable AI (XAI) refers to AI systems that are interpretable or understandable to humans. The research fields of AI ethics and XAI lack a common framework and conceptualization. There is no clarity of the field's depth and versatility. A systematic approach to understanding the corpus is needed. A systematic review offers an opportunity to detect research gaps and focus points. This article presents the results of a systematic mapping study (SMS) of the research field of the Ethics of AI. The focus is on understanding the role of XAI and how the topic has been studied empirically. An SMS is a tool for performing a repeatable and continuable literature search. This article contributes to the research field with a Systematic Map that visualizes what, how, when, and why XAI has been studied empirically in the field of AI ethics. The mapping reveals research gaps in the area. Empirical contributions are drawn from the analysis. The contributions are reflected on in regards to theoretical and practical implications. As the scope of the SMS is a broader research area of AI ethics, the collected dataset opens possibilities to continue the mapping process in other directions.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; *Philosophical/theoretical foundations of artificial intelligence*;

Additional Key Words and Phrases: AI ethics, explainable AI, artificial intelligence, systematic mapping study

ACM Reference format:
Heidi Vainio-Pekka, Mamia Ori-otse Agbese, Marianna Jantunen, Ville Vakkuri, Tommi Mikkonen, Rebekah Rousi, and Pekka Abrahamsson. 2023. The Role of Explainable AI in the Research Field of AI Ethics. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 26 (December 2023), 39 pages.
https://doi.org/10.1145/3599974

Authors' addresses: H. Vainio-Pekka, M. O.-O. Agbese, M. Jantunen, and T. Mikkonen, University of Jyväskylä, Mattilanniemi 2, Jyväskylä, Finland, 40100; emails: heidi.s.vainio-pekka@student.jyu.fi, {mamia.o.agbese, marianna.s.p.jantunen, tommi.j.mikkonen}@jyu.fi; V. Vakkuri and R. Rousi, University of Vaasa, Wolffintie 34, Vaasa, Finland; emails: {ville.vakkuri, rebekah.rousi}@uwasa.fi; P. Abrahamsson, Tampere University, Kalevantie 4, Tampere, Finland; email: pekka.abrahamsson@tuni.fi.

2160-6455/2023/12-ART26
https://doi.org/10.1145/3599974

1 INTRODUCTION

Artificial Intelligence (AI) is one of the most prominent and influential technologies of modern times. Its rapid development and increasing human dependency on it has facilitated the adoption of AI in almost all imaginable sectors of life [13]. Furthermore, AI's proliferation in critical areas, its speed of development, and the race between nations and companies to build robust AI tools has increased the need to set ethical guidelines and principles for AI development and deployment.

AI ethics is a burgeoning research field that has emerged in response to the challenges related to the impact of AI. The challenges posed by AI include data bias, privacy, and fairness issues, in addition to the requirement for AI practitioners to gain better knowledge about the impact of the technology. As such, the subject of AI ethics itself is versatile, ranging from highly technical issues to understanding human behavior in the research, interaction, development, and usage of AI [113]. AI ethics is often broken down into principles, such as transparency, responsibility, trust, privacy, sustainability, autonomy, and dignity. Five of these principles have emerged as dominant, including transparency, justice and fairness, non-maleficence, responsibility, and privacy [89]. Transparency, which is arguably the most prevalent [89], is often viewed as a pro-ethical principle and an enabler for ethical AI [166]. Consequently, transparency plays an important role in AI ethics, where it covers a broad scope that includes XAI [101]. XAI refers to an interpretable system that provides an understandable explanation of the system output [2]. XAI draws attention to the area of AI ethics research focused on how AI systems make decisions, the explanations of the decisions and how the decisions are communicated to relevant stakeholders [91].

XAI is a growing area of research, especially as AI systems are implemented in critical sectors that warrant transparency for AI actions. One example of this is medical AI, in which the need for an understandable system is tied to the core ethical values of medicine [11]. Here, expectations for explainability are high [83]. However, due to its novelty, the field remains riddled with unclarity and lack of structure. Despite its importance, the role of transparency is not well defined in AI ethics. Moreover, XAI currently suffers from a lack of commonly agreed definitions of core concepts [53, 89]. Most of the research and reviews of XAI in view of AI ethics are tailored toward a particular aspect of explainability, such as algorithm explanations [155, 191], black-box explanations [71], and methods that aim to describe explainability [179]. A recent systematic review [184] helped to explore current approaches and limitations for XAI. However, the review focuses on the area of reinforcement learning with no recourse to its role in AI ethics. Consequently, there is currently limited research that explores XAI and its specific part in AI ethics in depth.

Given the gap in previous studies, this article examines the research field of XAI and its role in AI ethics scholarship. The article's research question, "*What is the role of XAI in the AI ethics research field?*" requires an overview of the corpus of academic literature on AI ethics. The focus of the article is on concrete, actionable issues rather than philosophical discussion, with the main emphasis on empirical research studies.

The article adopts an SMS to map the research literature of AI ethics. SMS is a form of **Systematic Literature Review (SLR)** [93]. SLR and SMS are secondary studies where the attention is placed on analyzing the evidence of previous research. SLR aims to find and evaluate the relevant papers, which are called primary studies, on a specific research area. SMS aims to identify and categorize the existing literature more in general [93]. High-quality SMSs can have a significant benefit for the research area in establishing baselines for future research [93].

To understand the role of XAI in the research field of AI ethics, SMS methodology represents a better approach than SLR. The infancy and lack of coherence of the AI Ethics research area support the use of SMS. The size of the research area is unknown, and the role of XAI is new. The conceptual ambiguity of the research area [89] necessitates SMS usage. Several SMSs are studied, and guidelines are utilized. However, the most influential papers for this study are the guidelines

of Reference [130] and the SMS of Reference [128]. This article builds on the SMS of Vakkuri and Abrahamsson [168].

The rest of this article proceeds as follows: Section 2 serves as a background for XAI and related AI topics, machine learning, and the principles for ethical AI. Section 3 reports the literature search process. The section starts with a theoretical framework of SMS and continues with reporting the use of SMS in this article. The literature search process results in primary studies ($n = 142$) that form the scope of this study.

Section 4 presents the classification schema and the numeric results of classification. Section 5 presents the systematic mapping, where the results are analyzed and compared, and the annual trends and the publication venues are investigated. Section 6 proposes theoretical and practical implications of primary empirical contributions. Section 7 proposes some future research topics. Finally, towards the end of the article, Section 8 draws some final conclusions.

2 BACKGROUND

AI has a long history in software development, with its roots stretching back to the 1950s [111]. During its history, AI has had its ups and downs in the hype curve, making it appear brand-new at certain intervals in public discourse. Although there has been a lack of build-up around AI in the industrial sector, AI has been a standard part of the industrial repertoire ever since the 1980s [36]. However, it was not until 2007 that the introduction and generalization of smartphones and social media channels started to generate large amounts of data. This affected machine learning by providing it with training material and target applications [36].

As the most common form of AI today, machine learning has been coded to learn either by human supervision or independently with training data. By the definition of Reference [10], machine learning refers to a computer program that is programmed to optimize its performance using example data or past experience. Machine learning models can be used to make future predictions or gain knowledge from the past [10].

Data is the key to train a machine learning model. Although the amount of data is growing exponentially, the major challenge is the usability of the data. This is because raw data are unlabeled or unstructured and require extensive refinement efforts. Techniques like deep learning can be used as part of a solution, because deep learning requires a smaller training dataset. A deep learning model can be fed with raw data and used for detection and classification. Models using unsupervised deep learning are expected to become more critical in the future. Deep learning can be employed for more complex tasks such as natural language recognition and imitation of human vision, and in the future, it can be combined with complex reasoning [103].

2.1 AI Ethics

Because of the capability of AI systems to learn and make decisions autonomously, coupled by the broad concern in deploying AI in various fields, the interest and need for ethical research and guidelines have increased. In academia, discussions and research on AI ethics have been running for decades. Yet, these initiatives rarely cross with the development of AI systems [168]. Research on AI ethics has been focusing on the potential of AI on a theoretical level and on finding technological solutions. However, a broader perspective is often required [34]. AI ethics is a continually evolving research area that holds relevance for several domains, including computer science, economics, and philosophy. The research consists of a large variety of papers from different areas concerning AI ethics, which makes the definition of the field of AI ethics a challenging task [168].

The ethics of AI is often defined using lists of principles, laws, or guidelines for AI developers or implementers to follow [200]. Jobin et al. [89] mapped the corpus, including the grey literature (e.g., corporations' white papers and reports) of AI ethical guidelines and principles. The results

revealed five primary principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. The interpretation of these principles varies, depending on the domain, actors, and issue. Transparency is interpreted as explainability, understandability, interpretability, communication, disclosure, and showing. Justice is most often interpreted as fairness, consistency, inclusion, equality, equity, (non-)bias, and (non-)discrimination. Most frequently, non-maleficence refers to general security, safety, and not causing of foreseeable or unintentional harm. Responsibility and accountability refers to liability and integrity or to the different actors named as accountable for AI's actions. Finally, privacy in AI ethics is both a value to uphold and a right to be protected [89].

The most frequent requirement in AI ethics literature is transparency, followed by the requirements of justice and fairness [89]. Transparency is often needed to ensure the system's ethical functioning, because without transparency, fairness cannot be evidenced in the system. A third, closely connected issue is accountability. Together, these three elements construct the **fairness, accountability, and transparency (FAT)** theorem. In recent years, the questions about responsibility and transparency in autonomous systems have been raised in mainstream media due to pedestrian fatalities with self-driving cars. Autonomous driving is a broadly discussed topic in the AI ethics field. It has opened an avenue for non-practitioners to join the conversation and understand the issues related to AI ethics. Massachusetts Institute of Technology's "Moral Machine" research [16] collected 40 million answers to their online experiment, which studied decisions in ethical situations related to autonomous driving. In recent years, the discussion around AI ethics has opened to incorporate a broader scope.

Governments and regulators like the **European Union (EU)** are increasingly becoming interested in the topic of AI ethics. European Commission's AI High-level Expert Group [5] has identified "Trustworthy AI" as the EU's foundational ambition for ethical AI. Companies and private organizations are also establishing ethical frameworks and principles. Large practicing organizations, such as Google, Intel, and Microsoft, have also presented their guidelines concerning ethics in AI [169]. In academia, guidelines and principles aim to structure the research field. One notable example is the IEEE standard for Ethically Aligned Design [120].

Frameworks and guidelines may be a good starting point for the conversation, but they are not sufficient to solve the challenge of AI ethics without other measures in place. The challenge of frameworks is that they tend to lack practices and modeled behavior upon which to implement them. Furthermore, they often require more work to be production-ready [118]. Often, the principles and associated frameworks presented in the literature are not actively used in practice [171]. By the end of 2020, there were over 100 sets of principles, many of which were vaguely formulated [42]. Hence, choosing the right framework from all available ones may be a challenging decision, because AI ethics lacks the commonly agreed ethical framework [60]. Also, there is a lack of existing methodology in identifying the relevant frameworks for AI development in the context of implementing explainability [177]. The choice of suitable methods to create AI with the desired outcome extends beyond frameworks and must be made in each case individually, considering the needs of the relevant stakeholders and the desired explanation method properties [177].

One notable connection to AI ethics is the concept of Responsible AI, a paradigm to ensure that fairness, model explainability, and accountability are included in the practical implementation of AI methods. Besides AI principles, the Responsible AI practices include technical and non-technical training, guidance and tools to avoid and mitigate issues that may arise, and a governance model to assign responsibilities and accountabilities. Where there are many organizations that are listing their AI principles, there are viewer examples of how to implement the AI principles into practice. For practical implementation list of principles is not solely enough, but Responsible AI practices are required [14].

2.2 XAI

XAI refers to an AI system that can explain its decisions [146]. AI technologies such as machine and deep learning techniques are used for automating and optimizing predictive data patterns to achieve better or faster decision-making. However, the complexity of techniques such as deep learning makes the resulting decisions hard to understand for humans. Thus, explanations can help communicate the justification behind a decision or action. This can engender trust in the decision [80]. Such transparency can also ensure that the complexity of the explanation matches the complexity capacity of the consumer [80].

Understanding human decision-making and explanation definition provides good grounds for XAI that requires multidisciplinary collaboration and the use of existing research from social sciences, such as philosophy, psychology, and cognitive science [114]. Explainability is viewed as important in assigning responsibility in cases of a system failure [141], such as a collision incident of a self-driving car. To ensure the right for explanations, legislation such as the **General Data Protection Regulation (GDPR)** outlines individuals' right for a meaningful explanation of decisions made by automated systems. However, while calls for XAI have increased, there have also been some arguments against it. Some AI researchers have advocated that, since humans are unable to provide exact explanations for their decisions, AI systems should not be expected to do so either [56, 80].

Another aspect of XAI is interpretability. AI models are expected to be interpretable, which means that they can explain the decision in understandable terms to a human [82]. Interpretability deals with understanding the algorithm output to be implemented for end-users [62]. Sophisticated knowledge extraction and preference elicitation is required to extract a meaningful explanation from the raw data used in the decision-making process [146]. This often means that a tradeoff must be made between accuracy, effectiveness, and interpretability [2]. Interpretability is not merely a technical problem; to gain interpretability of machine learning systems, it is necessary to focus on humans rather than technical aspects and provide personalized explanations to individuals [146].

Interpretability may not be expected from AI systems when users trust the system, even if it is known to be imperfect, or when the consequences of a wrong decision are considered insignificant [82]. Interpretability has divergent requirements depending on the stakeholders involved [82]. Overall, interpretability requires explanations at varying degrees to help illuminate decisions made by AI [141]. Reasons behind the need for XAI vary. Based on Wachter et al. [180], the reasons may be as follows: (1) to inform the subject of the reasoning for a particular decision or explain the reasons for rejection; and (2) to understand how the decision-model needs to be changed to receive the desired decisions in the future. Overall, the application area and purpose may determine the need for interpretability.

Explainable and understandable systems are required for society to trust and accept algorithmic decision-making systems [180]. Better explanations can also improve existing models and open new opportunities, such as the use of machines for teaching humans [146]. XAI is also a potential tool to detect flaws in the system, decrease biases in the data, and gain new insights into the problem at hand [141], this can help ensure transparency of the system.

2.2.1 Transparency. The meaning of transparency varies depending on the subject. As a result, the concept is vague, making misinterpretations likely. In the discipline of information management, transparency often refers to the form of information visibility, such as access to information [166]. In computer science and IT disciplines, transparency often refers to a condition of information visibility, such as the transparency of a computer application to its users, as well as how much and what information is made accessible to a particular user by the information provider [166]. In this article, the term "transparency" is used in the sense of the condition of information visibility.

Although transparency is often required, it is not easy to provide. The information provider (e.g., company or public institution) must define who has the right to access the information and the accessibility conditions for it [166]. Legislation such as GDPR may control the access and sharing of a specific type of information between users.

As mentioned above, transparency is listed as one of the primary principles of AI ethics [89]. At the same time, transparency can actually be seen as the pro-ethical circumstance that makes the implementation of AI ethics possible in the first place. Without understanding how the system works, it is impossible to understand why it malfunctioned, and consequently, to establish who is accountable for the malfunction's effects. Instead of seeing transparency as an ethical principle, it would be more accurate to treat it as an ethically enabling or impairing factor, or as described above, a pro-ethical condition. Information transparency enables ethical implementation when the system provides the information necessary for the endorsement of ethical principles or when it provides details on how information is constrained. Transparency can impair ethical principles if it gives misinformation or inadequate information or exposes an excessive amount of information. The impairing of ethical principles could lead to challenges, for example, with discrimination, privacy, and security [166]. Transparency is normally associated with the black-box problem in AI ethics.

2.2.2 Black-box Problem. The term "black box" is used when the AI model is not understandable and cannot provide a suitable explanation for its decisions [2]. A black box refers to a model that is either too complicated for any human to comprehend or proprietary to someone [139]. To understand the black box, the model needs to be built to be interpretable, or a second model must be created that explains the first black-box model [139]. Interpretability in the AI context refers to the capability to understand the overall operational logic in machine learning algorithms, not just the answer [2]. The terms *interpretability* and *explainability* are often used as synonyms [2], but this can be challenging, because there is a subtle difference between them related to the level of required understandability. In public discussions, the term "Explainable" AI is more often referred to than "Interpretable" AI, whereas, in academic discourse, the situation is contrary [2]. Current AI regulation, such as GDPR, requires the right to explanation, not an interpretable model, which might cause problems, as only requiring an explanation does not require the explanation to be accurate and/or complete, and therefore right for explanation is an incomplete requirement [139].

A second post hoc explainable model may provide explanations that do not make sense or that are not detailed enough to understand in terms of what the black box is doing. To acquire a full understanding of the model, the information provided by its transparency should also be interpretable. Secondary explanatory models are often incompatible with information outside the black box. The lack of transparency in the whole decision process may prevent interpretation by human decision-makers. Secondary models can also lead to overly complicated decision pathways when transparency is actually required from two models (i.e., the original black box and the explanatory model) [139].

Neither interpretable machine learning model is challenge-free. First, this is because it is a computational challenge to build such a model. Second, the AI system's total transparency can jeopardize the system owner's business logic, because the system owner must give away intellectual property [45]. In addition, constructing an interpretable model is often expensive, because this requires domain-specific knowledge, and there are no general solutions that would work in different use cases. In creating an interpretable model, it is a challenge to find the balance between interpretability and accuracy, because interpretable models tend to reveal hidden patterns in data that are not relevant to the subject [139, 140].

2.2.3 Accountability and Algorithmic Bias. In addition to interpretable machine learning and black-box problems, core concepts around XAI include AI's accuracy, a performance metric to compare the number of correct predictions to all predictions, and responsible AI [2]. Accountability refers to an actor who is accountable for the decisions made by AI. To establish accountability, the system must be understandable. A lack of transparency and accountability in predictive models can cause serious problems, such as discrimination in the juridical system, endangering a person's health, or misuse of valuable resources [171]. Based on Vakkuri's [171] research, transparency is the enabler for accountability, and together, transparency and accountability motivate responsibility. Finally, responsibility produces fairness. Fairness is often linked with algorithmic biases. In other words, an AI system might repeat and magnify biases in our society, such as by segregating groups with a history of being marginalized (e.g., in preferring men over women or discriminating against people of color).

Machine learning bias is defined as "any basis for choosing one generalization over another, other than strict consistency with the instances" [117]. Machine learning systems are neutral and do not have opinions, but the models are not used in voids, which makes them vulnerable to human bias. In the context of machine learning models, discrimination and unfairness in the models can be caused by unfairness in the data and the collection and processing of data or the selected machine learning system. The practical deployment of the system may reveal biases that were invisible during the development process. Ultimately, there is no easy solution to ensure fairness of algorithmic decisions [175]. But, there is an interest in finding a working solution.

Veale and Binns [175] identified three distinctive approaches to ensure fairer machine learning. The first is the third-party approach, where an outside organization manages data fairness for the main organization. The second is the collaborative knowledge base approach, where linked databases containing fairness issues are flagged by researchers and practitioners. Finally, the third approach is an exploratory approach, where exploratory fairness analysis of the data is performed before training or practically implementing the model. In this article, the interest is in the exploratory approach, because it is connected to the black-box problem [175]. The biases are studied from the perspective of XAI, which aims to bring transparency to the AI system. Less emphasis is dedicated to research on how data can be collected or processed to avoid biases.

2.3 Summary of Emerging Issues

AI ethics research lacks harmony and standard agreement on defining the core principles of the field [45, 89]. Moreover, the research field of XAI is complex and is in need of a common vocabulary and formalization [54]. This article aims not to solve the issue of definitions of fairness and transparency but rather to investigate the existing research connected to transparency as understood in this article, as a requirement for the AI system to provide an understandable explanation if needed in the context of the application. This requirement applies to systems that are non-explainable because of the training method or biased as a result of bias in the training data. This article takes no stand upon ranking the principles. Instead, it aims to provide a more in-depth understanding of what has been studied and how in terms of transparent and explainable AI systems.

The research field of XAI studied as a sub-field of AI ethics examines the challenges and looks for potential solutions for transparent machine learning models, aiming to enable the fulfillment of such ethical principles as accountability, responsibility, and fairness [157]. XAI can benefit a broad range of domains relying on AI systems. Especially in domains such as law, finance, military, and transportation, the need for XAI is emphasized [2]. In such areas, AI systems have a direct influence on the physical conditions of people and can cause injuries [2]. In other domains, transparency may not be a critical requirement. There is no one-for-all framework or solution available for transparency issues. Hence, domain-specific solutions and frameworks are required.

Adadi et al.'s [2] research showed that the impact of XAI is spanning a broad range of application domains. However, the lack of formalism regarding problem formulation, divergence in explanation methods and results [95], and clear unambiguous definitions burdens the research field. Moreover, they noted that the human's role is not sufficiently studied [45, 56]. A recently published paper recognized the same challenge with the lack of user-centric design in XAI [58]. For implementation, it is important to understand user requirements and needs to ensure trust and acceptance of algorithmic decision systems [155]. In addition to understanding the user's needs, the research field lacks knowledge on industrial practices with AI ethics [171] and knowledge on how different explanation methods result in varied results. Overall, there is a concern that the XAI field suffers from the distancing of real-world problems [139].

AI ethics and XAI are broad, versatile topics with increasing importance. The present SMS is timely, as it enables an understanding of what has been studied in AI ethics. It is required to understand what is studied in AI ethics research to clarify the role of explainable AI. More systematic research is required for this purpose, and in the next sections, an SMS is used to understand the study field of AI ethics and how XAI is manifested in the research.

3 LITERATURE SEARCH FOR PRIMARY STUDIES

This study employed the SMS method. The main focus of SMS is to "provide an overview of a research area, and identify the quantity and type of research and results available within it" [130]. The SMS aims to identify the potential research gaps and trends, including the understudied topics and research types. The expected outcome for SMS is to identify and choose the primary studies and map the literature [93].

The research builds on an SMS developed by Vakkuri and Abrahamsson [168], who studied the key concepts in the field of AI ethics. For this article, the research was updated twice, first during mid-way through 2020 and later in the last quarter of 2021. In this article, the goal is to analyze how XAI is researched in the study field of AI ethics. The interest is in practical implementation and connection to real-world issues. Thus, the focus is on empirical studies, and papers without data analysis, such as literature reviews, were considered theoretical. We included papers analyzing empirical data regardless of the data type, or data collection or analysis method.

The research question for an SMS can cover issues such as what topics are addressed, what empirical methods are used, and what sub-topics have been sufficiently empirically studied [93]. This guideline forms the basis of the current research question, "What is the role of explainable AI in the AI ethics research field?" and its three sub-questions:

R1 What has been empirically researched in the field of AI ethics?
R2 What is the state of published research on XAI in the field of AI ethics in the past 10 years?
R3 Where are the research gaps in the field?

To answer the main research question, it is first important to answer the first sub-question [R1]. In this article, the question is studied on a superficial level to offer enough background to understand the main research question. The major topics are noted, and the research field's size and proportion of empirical research from the existing academic literature are delineated.

To address the second question [R2] and to understand XAI's role and importance in AI ethics, research with XAI as the focus is reflected against a full dataset of empirical studies. More in-depth analysis and classification are performed on papers focusing on XAI to understand what, how, and why it has been studied in past 10 years. The analysis includes investigation of research methods, contributions, focus, and pertinence to XAI. In addition, the annual changes in the research field are studied to reveal trends. The connection to real-world issues is also reviewed. This article

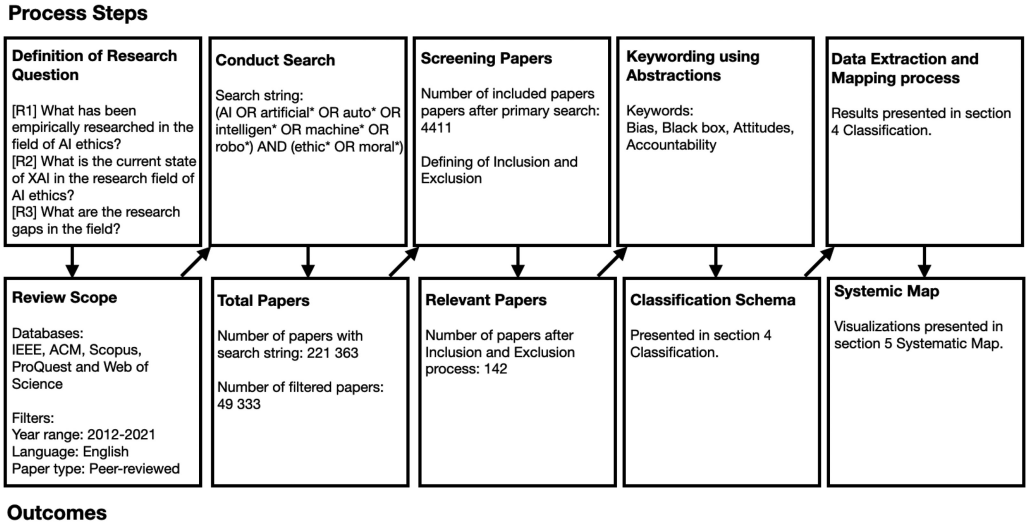

Fig. 1. SMS process based on Petersen et al. (2018).

investigates the current research corpus with empirical evidence to understand the AI ethics research field in a way that is closer to real-world issues.

The third question [R3] can be addressed based on a background literature review and a profound SMS. The background literature review revealed gaps, such as the lack of understanding of the human role in XAI [2] that were also highlighted in SMS analysis.

The processes of building an SMS are cumulative, and it includes several rounds of screening papers. The process steps and outcomes are presented in Figure 1 based on Reference [130]. The headline of each block describes the process step, and the body reflects this study. The figure guides the reader through the entire study.

Due to the fact that an SMS's goal is to understand the research area rather than give evidence, the articles do not need to feature in-depth examination. Thus, the number of articles included can be larger [130]. The total number of papers included from five databases, after deleting duplicates, was 4,411. After applying the inclusion and exclusion criteria, the sample was narrowed to 142 papers. In the following, each step is further explained based on the theoretical framework.

3.1 Primary Search

The first step in an SMS is to identify the primary studies that contain relevant research results [37]. This article builds on the SMS of Vakkuri and Abrahamsson [168], and the search strings and selected databases were adopted from their research. With the research question of, "What topics are covered in AI ethics research?" the search string consisted of the two following parts: (1) AI and its synonyms (robotics, artificial, intelligence, machine, and autonomous); and (2) ethics and its synonyms (morals). The search string was as follows:

```
(AI OR artificial* OR auto* OR intelligen* OR machine* OR robo*) AND
(ethic* OR moral*)
```

The search was narrowed to include only the headline and abstract. The search was performed in the five following electronic databases: IEEE, ACM, Scopus, ProQuest, and Web of Science. In total, there were 221,363 results. Table 1 shows the results of primary search per database.

Table 1. Results of Primary Search: 2012–2021

Database	Total papers	Filtered papers	Selected papers
IEEE Xplore	5,132	2,437	861
ACM Digital Library	1,121	914	739
Scopus	58,081	19,822	3,326
ProQuest	132,410	13,457	1,038
Web of Science	24,619	12,703	1,084
Total	221,363	49,333	7,048

INCLUSION AND EXCLUSION CRITERIA

INCLUSION	EXCLUSION
[I1] Papers focused on Ethics of AI	[E1] Papers only mentioning AI ethics in general introduction
[I2] Year range: 2012-2020	[E2] Papers with empirical research
[I3] Peer reviewed articles and proceeding papers	[E3] Papers not related to XAI or transparency
[I4] Language: English	
[I5] Full access	
[I6] White literature	

Fig. 2. Inclusion and exclusion criteria.

Because of rapid progress in the development of AI in early 2010s, previous studies, such as those carried out before 2012, are often not as relevant as the more current research. Thus, these were excluded from the results. Since the aim is to understand the state of academic research related to the topic, only peer-reviewed articles were included [20]. The search with four filters (document type, publication year, peer-reviewed, and language) performed in five databases resulted in 49,333 papers. All the abstracts of the resulted papers were screened manually to exclude papers that were irrelevant to the study. The primary search was done first in 2016 and updated in 2019 and 2021. Manual screening was executed by the four first authors. At this stage, each paper was screened once. To guarantee consistency between readers, if the reader was uncertain, then the paper was included. The primary search resulted in 7,048 papers, which were combined into one dataset, and duplicates were deleted. The remaining papers amounted to 4,411 that were left for closer review in the inclusion and exclusion process.

3.2 Inclusion and Exclusion

The second step of SMS is to examine the selected papers and find the primary studies [37]. This process requires defining a greater number of narrower inclusion criteria. The inclusion process is guided by the research goal and desirable contribution [128]. The inclusion and exclusion criteria are presented in Figure 2.

The study's aim is to map the relevant research area of the ethics of AI in the domain of information system science. Hence, in this step, only papers focusing on the ethics of AI [I1] were included. Because many papers were included after the primary search, it was decided to include only the papers with full access [I5]. The inclusion criteria from the primary search (year range [I2], academic peer-reviewed papers [I3], and language [I4]) were cross-checked during the inclusion process. To guarantee the high academic quality of the included papers, only white literature,

Table 2. Excluded Papers

Reason for exclusion	Number of papers
Duplicate	2,637
Inadequate Academic Quality [I3, I6]	391
Not Fully Available [I5]	534
Language [I4]	21
Out of Scope [I1, E1]	1,279
Theoretical - No empirical data used [E2]	1,653
Not related to XAI [E3]	361

peer-reviewed, papers were included [I6]. White literature refers to full papers published in venues of high control and credibility, and it excludes pre-prints, technical reports, blogs, and other types of publications that are referred to as grey and black literature [65].

In SMS studies, exclusion criteria may require excluding papers that only mention the main interest area in the abstract. General concepts are often used in abstracts, even if the paper focuses on something else [130]. The first exclusion criterion [E1] is the exclusion of papers that do not contribute to AI ethics research and only mention the potential ethical issues related to AI in the general introduction. Moreover, in this article, the interest is in practical AI implementation rather than a philosophical concern. Therefore, papers without empirical research were excluded from the study [E2]. In the final screening, papers that did not focus on XAI or related topics were excluded [E3].

The inclusion and exclusion criteria were established and defined during the screening process. The inclusion criteria provided the general boundary and quality conditions, and the exclusion criteria gave more detailed limitations to distinguish the sample relevant for this article.

For the first screening round, three quality inclusion rules were applied: language [I4]; access to full text [I5]; and sufficiently used references as well as overall academic quality [I6]. This means that workshop, keynote, panel, and paper presentations were excluded, along with short papers, tutorials, and abstracts. In addition, papers that did not focus on the ethics of AI were excluded [E1]. During the screening round, the quality of each paper was validated. Papers that did not meet the academic peer-review standards, such as short papers, tutorials, and panel/keynote/workshop presentations, were excluded from the study.

The included papers were clustered into two categories, theoretical and empirical, to separate the empirical papers that were meaningful for this article's goal. The empirical papers were manually separated during the screening, because this was considered the most reliable way to ensure the sample would include all the relevant papers. The screening was executed by the first four authors. Each paper was screened by one or two authors. If the first reader was uncertain, then the second opinion was provided. From the total of 2,192 papers that met the inclusion criteria, 503 used empirical material. The theoretical papers consist of reports, opinions, philosophical papers, problem descriptions, proposals, and academic literature reviews.

For the second screening, the papers were skimmed and scanned for keywords based on the focus area to find the papers connected to XAI. As described in the second section, XAI is a vague concept, and there is no commonly agreed framework on what topics should be included under the term. Thus, papers focusing on responsible AI, algorithmic bias, or black-box models were included to ensure the inclusion of all relevant papers. The excluded papers are visualized in Table 2.

The primary studies ($n = 142$) included in the SMS are further classified and analyzed in the next section. The full sample of papers with empirical evidence ($n = 503$) was further reviewed to

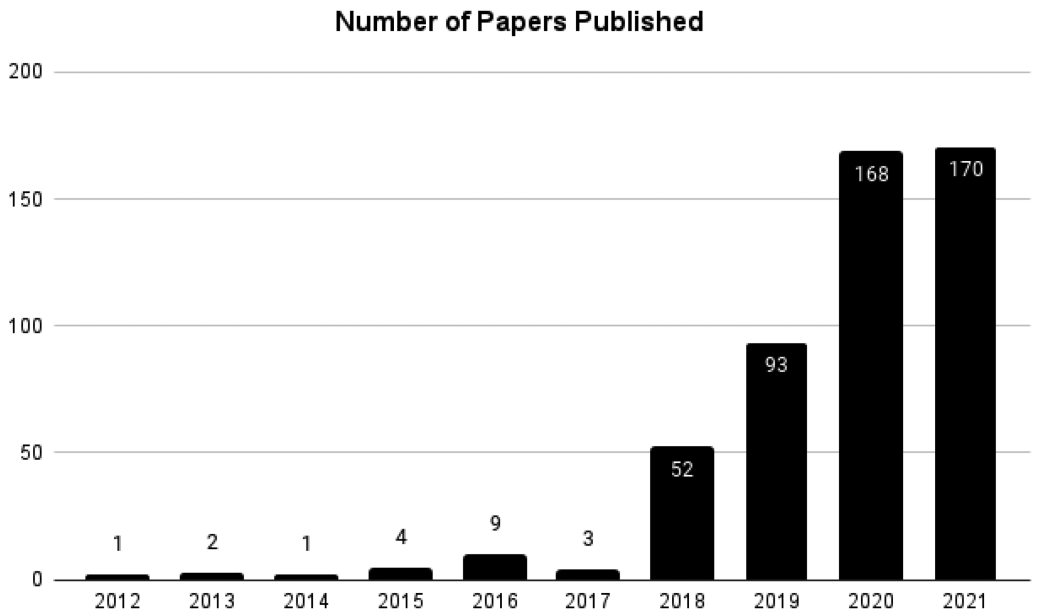

Fig. 3. Annual changes in publication of empirical papers in AI ethics research area.

understand the overall field of AI ethics described in them. However, the analysis was done on a superficial level, because a more thorough investigation was outside the scope of this study.

3.3 Short Analysis of AI Ethics Research Field with Empirical Evidence

Future studies are required to understand the research area of AI ethics more comprehensively. Yet, this short analysis gives sufficient background to reflect the role of XAI against the full sample of AI ethics research with empirical evidence (n = 503). The empirical papers represent 23% of the whole sample of manually included papers (n = 2,192). This finding forms the first **empirical contribution (EC)**.

- EC1: Most of the research papers in the field of AI ethics do not use empirical evidence. Only 23% of the papers provide empirical evidence.

The two following dimensions were observed within the entire sample: emerging themes and the year of publication. The theme analysis was done during the keywording process described in the next section. A more profound analysis would require a more systematic approach.

Since the research area is in its infancy, the year of publication can provide insight into the research area's growth. The papers published per year are visualized in Figure 3. The size of the bar presents the number of papers published each year.

The visualization reveals significant growth starting from 2018. There is a clear correlation to public discussions, with discourse on AI ethics growing significantly in media in 2018 [124]. This finding forms the second empirical contribution.

- EC2: Empirical research on AI ethics grew significantly in 2018, corresponding with trends in public discourse.

Based on the shallow categorization of the topics during the classification, most papers focused on general issues and challenges related to AI ethics. Some notable topics in the research field were human-robot interaction for both physical and virtual robots (focus in 77 of 503 papers),

autonomous vehicles (58 of 503 papers), health and care (54 of 503 papers), education (31 of 503 papers), and governance/regulation (28 of 503 papers). The papers related to XAI (n = 142) represent 28% of the full sample of empirical papers (n = 503). This finding forms the first **primary empirical contribution (PEC)**.

- PEC1: XAI is a significant research focus in the study field of AI Ethics. Of the empirical research papers published after 2012, 28% are related to XAI.

Since the inclusion of XAI did not require the paper to have full dedication and focus on XAI, the number of papers engaging with XAI is not comparable to other emerging themes. In addition, papers with partial and marginal input to XAI were included if they contributed to the topic. No further examination was performed on excluded papers.

4 CLASSIFICATION

Classification uses a systematic process where the classification schema evolves and is specified during the process [130]. The first step, keywording, reduces the time required for building the classification schema and ensures that the classification schema represents existing studies [130]. The process was initiated during the last stage of the inclusion process and continued with the final sample, the primary studies, (n = 142) during the classification. Next, the classification schema, classification results, and the overview of the primary studies are presented.

4.1 Classification Schema

For the classification schema, the papers were examined in terms of the four facets adopted from SMS of Paternoster et al. [128]. These facets were research, contribution, focus, and pertinence.

(1) Research facet. The research type is used to distinguish between different types of studies and chosen research methodology. A research type *proposal of a solution* refers to papers proposing a novel solution technique and arguing for its relevance without full justification. At best, such papers provide a narrow proof of concept. *Validation research* papers investigate the properties of their or others' proposals of solutions that are not implemented in practice. The investigation is performed in a methodologically sound research setup. *Philosophical papers* propose new conceptual frameworks and structures. Finally, *experience papers* describe the implementation in practice, such as listing the lessons learned. The experience may be the author's or that of the person studied [185].

(2) Contribution facet. The aim is to identify the tangible contribution of the paper. This can be an operational *procedure* for development or analysis to provide a a new and more effective way to do something, such as a design framework. Alternatively, it can be a *model* representing the observed reality and structuring the problem area, an implemented computational *tool* to solve a particular problem, or a *specific solution* for a specific application problem. The contribution can also be a piece of generic *advice* with a less systematic approach than the model. It often focuses on one example case and is more vaguely directive than the procedure is. The contribution facet is based on Shawn's [153] research.

(3) Focus facet. Keywording that was performed during the last screening round revealed focus themes that were highlighted during the classification process. The focus themes detected were *algorithmic bias*, or the challenges with fairness because of biased and discriminative training data or model; *black box*, or the challenges with non-transparent systems; and *accountability*, with papers studying when and how the accountability of a non-transparent system is divided. Some papers focused on understanding the attitudes, expectations, and trust toward non-transparent systems. These papers were categorized as *attitude*.

Table 3. Results of Classification

Research Facet	N	Percentage
Proposal	83	58.5
Philosophical	24	16.9
Experience	21	14.8
Validation	14	9.9
Contribution Facet	N	Percentage
Tool	44	31.0
Model	36	25.4
Procedure	31	21.8
Specific Solution	18	12.7
Advice	13	9.2
Focus Facet	N	Percentage
Bias	65	45.8
Attitudes	40	28.2
Black Box	31	21.8
Accountability	6	4.2
Pertinence Facet	N	Percentage
Full	62	43.7
Partial	59	41.5
Marginal	21	14.8

(4) Pertinence facet. The pertinence facet shows the level of relation to XAI, which is the research focus of this article. The levels are as follows: *full*, where XAI or transparency issues are the main focus of the paper; *partial*, where the paper is partially related to XAI or transparency; and *marginal*, where the paper's primary research focus is out of transparency or XAI themes.

In all facets, the same paper can fit into several categories. Here, in such situations, the best possible fit was chosen. The process was highly opinion-based, and the evaluation of one individual could impair the study's quality and liability. Classification was done by the first author, and the classification schema was presented and evaluated by two reviewers to ensure the research quality.

4.2 Results of Classification

After the classification schema was established, the actual data extraction took place, and the articles were sorted into different classes. A significant portion of papers focused on biased algorithms. These papers were classified under the pertinence facet as "full" if the papers focused on making the whole system more transparent. Papers that focused on cleaning and fixing biased datasets were classified as having a "partial" pertinence toward XAI. They were considered to have a main focus that related more to data science. The pertinence facet helped clarify whether the paper has a strong focus on XAI and transparency issues. Papers with a marginal focus on XAI were seen to contribute to the topic even if the main focus was elsewhere, and therefore, they were kept in the sample.

After the classification, the papers were calculated in their respective classes and visualized with the number of papers in each facet's class and the percentage of the class compared with the full sample ($n = 142$). This highlights what has been emphasized in past research, revealing potential research gaps and possibilities for future research [130]. The classification results are presented in Table 3.

In the research facet, the proposal class was significantly emphasized, with 59% of the studies proposing a technical, mathematical, or design solution. The main contribution classes were tools

(computational solutions to a particular problem) and models (structuring the problem area). Many papers proposing a new computational tool suggested a new algorithm or mathematical solution.

Many papers focused on biased algorithms (46%). Papers where the main focus was to understand developers' and users' expectations, attitudes, and trust toward XAI systems represented 28% of the whole sample. From the attitudes category, only 14 papers (10% of the sample) focused on practitioners' expectations and opinions, and the remaining 26 papers focused on understanding how the general public sees the issue.

In addition to classification, the papers were clustered based on the publication venue (journal/conference) and type of data used (real-life or synthetic). Most papers, representing 99 papers (69%), were published at conferences. Only 10 papers (7%) used synthetic data, which indicates that the research on XAI is closely connected to real-life issues.

The overview of the primary studies ($n = 142$) in light of the the classification results is presented in Appendices. All the papers are found in the reference list at the end of this article. In the next section, the classified data are analyzed and visualized. The analysis aims to elucidate the study field of XAI and its role in AI ethics research.

5 SYSTEMATIC MAP

There are several ways to visualize the results of an SMS. The two most common approaches are bar plots and bubble plots [131]. Bubble plot visualization is exceptionally well-suited to illustrating the number of studies for a combination of categorizations [131]. Because the classification schema applied in this study includes several categories, the bubble diagrams were built to visualize the number of papers in different classes and investigate correlations between them. Since there were four main facets in the classification schema, it was necessary to create several diagrams to avoid over-complicating the view. Different types of visualizations were constructed based on the area of inspection. In the next sections, the results of the classification schema, pertinence, impact, annual change, and the venue of the study field are visualized and analyzed.

5.1 Systematic Map in the Bubble Plot Visualization

A bubble plot diagram helps to give a quick overview of the research field and support the analysis more effectively than the frequency tables [130]. Here, the bubble plot diagram was built using summary statistics presented in the previous section (Table 3). The diagram visualizes the frequencies and correlations between categories and facets. The bubble plot diagram comprises two x-y scatterplots with bubbles in category intersections. The same idea is used twice, on opposite sides of the same diagram, to show the intersection with the third facet on the x-axis [130].

In the first bubble plot, the contribution and research facets are compared to the focus facet. The size of a bubble indicates the number of papers that are at the intersection of the coordinates. Next to a bubble, there is the percentage of the total amount ($n = 142$) in the represented category of the x-axis. The bubble plot is presented in Figure 4.

The bubble plot diagram shows the emphasis on focus facets in each of the research and contribution facets. The bubble plot reveals that the most significant emphasis of the research facet is in proposals solving algorithmic biases.

- EC3: The most popular paper type in the research facet is a proposal for solving algorithmic bias.

In addition, the proposals for black-box issues are highlighted. Proposal research studies new and novel techniques to solve a particular issue. When compared to validation research, which studies a specific solution that has already been implemented in practice, the size of the proposals bubble is much larger, which indicates the research field's freshness. It may be that there are few

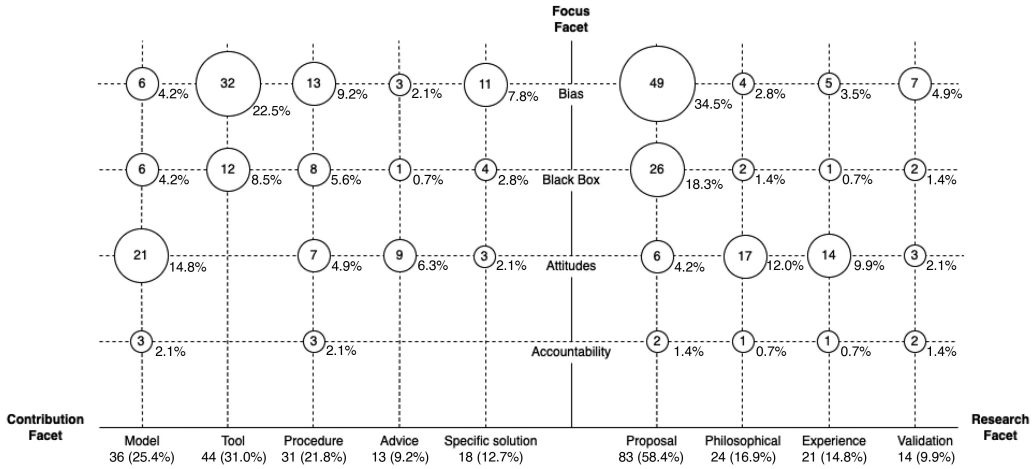

Fig. 4. Visualization of a systematic map in the form of a bubble plot.

proper practical solutions to fix the ethical issues related to XAI, these solutions are not yet implemented in practice, or the practical implementation has not yet been studied. The scarcity in the validation research is probably partly due to all the reasons mentioned above.

- PEC2: In the study field of Ethical XAI, the most common type of empirical research is studying a novel technique that can solve a computational challenge.

From the contribution facet, the largest bubble can be found at the intersection of bias and tools. Nearly one-fourth, 23% (32 papers), of the whole sample contributes to the research field with a computational solution to solve algorithmic biases. A computational tool to solve black-box issues was proposed in 12 papers.

- EC4: Almost one-quarter of the papers in the sample contribute to the research field with a computational solution to solve algorithmic biases.

In the contribution facet, the second-largest bubble (21 papers) can be found at the intersection of the attitude facet and the model facet. The bubble visualizes how the research field is modeled and structured by providing a better understanding of users and practitioners. Procedures, contributing by proposing a new way to solve such issues as design frameworks, are equally interesting in each focus facet when compared with the amount papers categorized per focus facet.

- EC5: Half of papers interested in users' and practitioners' attitudes and perceptions related to XAI and AI ethics are contributing by modeling and structuring the research area.

There is no strong weighting on any of the contribution types in the black box's focus facet. In the bias category there is an apparent weighting in the contribution of computational tools, and in the attitudes category there is weighting placed on modeling the problem area. From 32 papers that focus on bias and contribute with a computational tool, 30 papers (20% of the whole sample) have a research facet proposal. This is the most prevalent type of paper in the present study.

- EC6: The most prevalent paper type is that of a computational tool proposing a solution to a problem with bias. Every fifth paper presents this type of research.

From the bubble plot visualization, it can be concluded that the most common type of paper is a computational tool proposing to solve problems with biases, and in general, most papers look

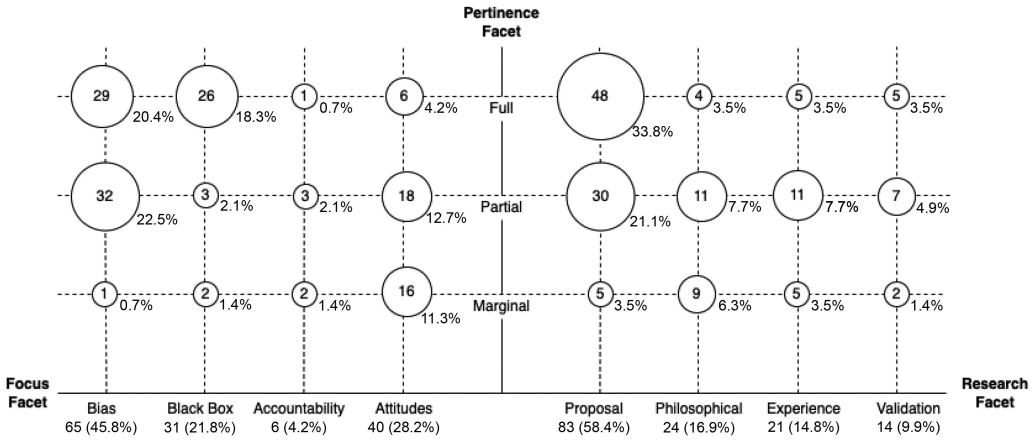

Fig. 5. Pertinence of the focus and research facets.

for novel techniques and solutions to computational problems. The results may indicate that the focus is slightly monotonous. Papers concerning black boxes, accountability, or attitudes are more dispersed, with the exception of the strong emphasis on proposals as a research type in the black-box papers. In addition, the results indicate immaturity in the research field.

- EC7: The research field seems to be somewhat monotonous and immature when considering the variety of topics, research methods used, and contributions of the papers.

5.2 Pertinence Mapped in a Bubble Plot

Since the pertinence indicates the accuracy in the XAI topic, pertinence was visualized with a bubble plot corresponding to the focus and research facets. The bubble plot visualization in Figure 5 aims to understand which focus areas and types of research have full pertinence on XAI and transparency-related topics and in which focus areas the pertinence remains elsewhere.

Out of the papers focusing on algorithmic biases, 44% had full focus in XAI. Many of the papers with partial focus had the main emphasis on cleaning data and fixing the datasets that are causing the discriminating and unfair decisions. These papers were considered to have their main pertinence in data science and fairness rather than in XAI. Not surprisingly, most of the papers (26 out of 31 papers) focusing on the black box were categorized to have full focus on XAI. The black box is one of the core concepts in XAI research [2].

- EC8: From the papers focusing on black box ($n = 31$) 84% had full pertinence on XAI.

In the results, 43% of papers with full pertinence were proposals of a novel solution. This again reflects the freshness of the research field, and it may indicate that the research done in the field is solution-oriented.

- EC9: The research field of XAI seems to be solution-oriented, and the research corpus with empirical evidence focuses more on finding solutions than exploring challenges.

Interestingly, only six papers with the main focus on attitudes and expectations of practitioners, users, and the public had full pertinence toward XAI. The results indicate a research gap in understanding people's perceptions of the topic. Similarly, papers with experience as a research focus had mainly partial or marginal focus on XAI.

- PEC3: The human perspective toward XAI is not well known. There is no in-depth understanding of the practitioners' and users' expectations and attitudes toward XAI.

It could be assumed that in research on XAI in AI ethics, there is a lack of understanding of the issues related to users' and practitioners' attitudes. Only four papers [41, 170, 171, 182] studied the current state of industrial implementation of AI ethics in general, and none of these had full pertinence to XAI. No paper studied the managerial or business perspectives of XAI.

- PEC4: Industrial implementation of XAI is not yet profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of XAI.

In total, only six papers presented a main focus on accountability. Although accountability was mentioned in several papers, it is interesting that it has not been more profoundly studied. Only one of the papers [138] had full pertinence toward XAI, and the rest related to AI ethics more generally. There seems to be a research gap in terms of understanding who takes responsibility and how this is decided if biased or non-transparent systems are not working as expected.

- EC10: There is a research gap in understanding who is responsible for the actions of non-transparent systems and how the responsibility is decided and communicated.

In conclusion, the pertinence was strongest in black-box research, and it was strongly present in the bias category. The attitudes category had a relatively weak connection to XAI. This indicates a need to understand better how people, including practitioners, businesses, and the public, perceive XAI.

5.3 Visualization of Annual Changes in the Research Field

The year range for the SMS described in this article was 2012–2021, but none of the papers from 2012–2016 were included in the study after the inclusion and exclusion processes were implemented. One paper from 2017, 16 papers from 2018, 40 papers from 2019, 52 papers 2020, and 33 papers from 2021 were included. Notably, the primary search was performed during September to December 2021. Hence, the record for 2021 is incomplete.

- EC11: XAI is a young but growing empirical research area in the field of AI ethics.

The growth of the research area seems to be stabilizing. From empirical papers (visualized in Section 3.3, Figure 3) published in 2019 ($n = 93$), 43% were connected to XAI; among those published in 2020 ($n = 167$), 31% (52 papers) were connected to XAI. Among empirical papers published in 2021 ($n = 170$), only 19% were related to XAI. This could be due to the faster growth of other research interests in the field of AI ethics or separation related to individual research agendas that were not so tightly connected to AI ethics. However, this study is only focused on XAI papers that are related to the research interest of AI ethics.

- EC12: The research interest in XAI compared with all published empirical papers on AI ethics was highest in 2019. Since then, the interest in XAI has grown yearly but not as rapidly as the empirical research on AI ethics has in general.

To visualize the annual changes in the research field, Figure 6 shows the annual changes and evolution in the contribution and research facets. The motivation for generating bubble plots was to detect trends in the research field. Although, as the research field is still emerging, the trends might only be seasonal changes. Moreover, because the year 2021 cannot be evaluated entirely, the results per year are not fully comparable.

The bubble plot reveals that the proposal has been the most popular category from the research facet every year. Experience and validation papers seem to be growing in popularity as the research field matures. Simultaneously, the number of philosophical papers is decreasing. The

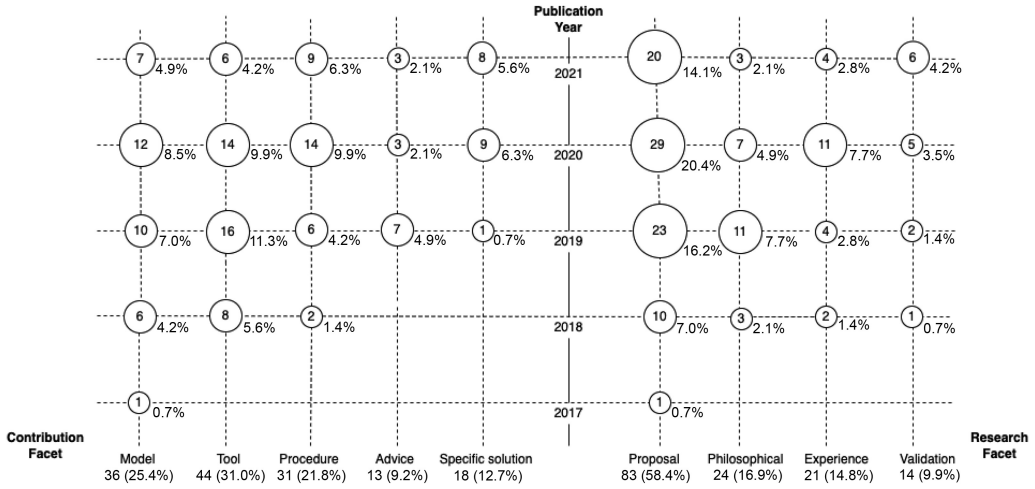

Fig. 6. Annual changes in the research and contribution facets.

research trend seems to be toward more practical understandings and less philosophical framing, as well as structuring of the focus area.

- EC13: The trend is toward more practical implications and less philosophical framing of the focus area.

In the contribution facet, the division between categories is more even. The strongest growth is in procedures, which are proposals for better ways of doing something. Interestingly, discussions on tools and computational solutions showed a decreasing trend in 2020 and 2021. This could indicate that the research field is evolving to become more holistic and not as intensely focused on finding technical solutions. Moreover, the growth in specific solutions could indicate that the computational tools are proposed to fix specific application issues. However, more research is required to verify this conclusion.

- EC14: The research contribution and interest seems to be shifting from proposing general computational solutions to proposing more holistic design/framework-level solutions and tools for specific application issues.

Another interesting observation is that advice papers seem to be decreasing in prevalence as the research field is maturing. This might be connected to the same trend to move from general advises to more application- or problem-specific solutions.

5.4 Venue and Focus of the Research

The research venue was studied to understand the quality and depth of the research area. All the papers were published either in conferences or journals. The papers published in journals should include the most mature research [86]. In addition, a higher degree of empirical evidence is expected from papers published in journals than from the conference of workshop proceedings [86].

As mentioned above, most papers were conference proceedings, representing 99 papers (69.7%). The most popular venue was the AAAI/ACM Conference on **AI, Ethics, and Society (AIES)**. Thirty-nine papers (28%) of the total sample (n = 142) were published in AIES.

- EC15: The most popular publication venue is AIES, with 28% of papers published in it.

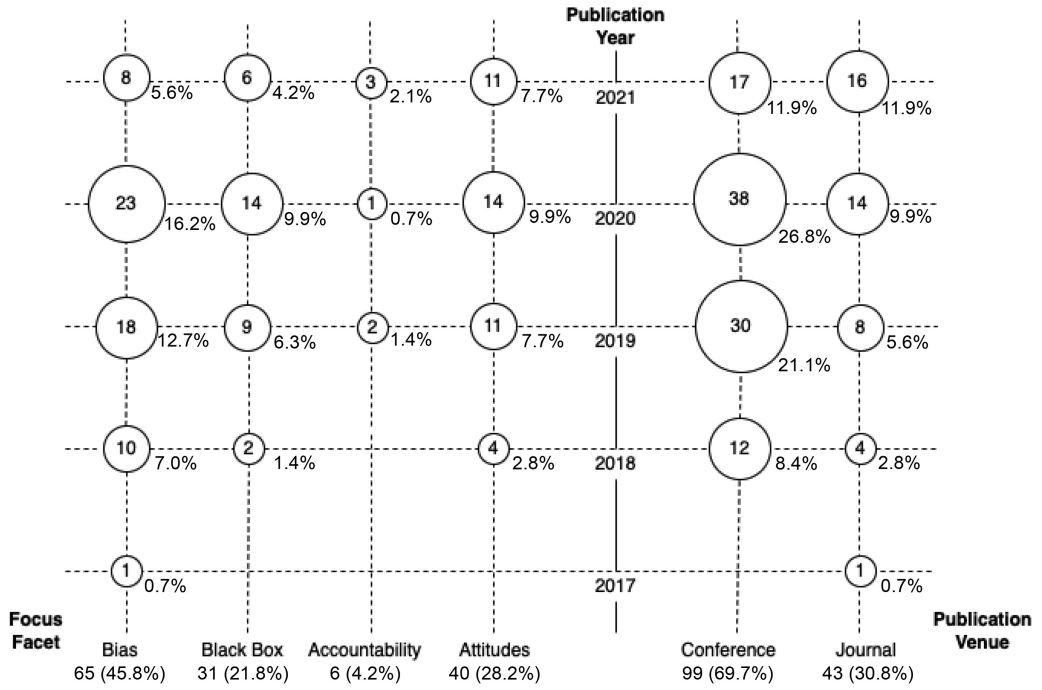

Fig. 7. Annual changes in publication venue and focus facet.

The annual variation of the publication venue and focus facet is visualized in Figure 7 with a view to elucidating how the research area has been evolving.

Interestingly, in 2021 almost as many papers were published in journals as in conferences, but, since the primary search was performed during late 2021, the incomplete nature of the data may have affected the result. The division between conference proceedings and journals since 2020 seems as expected, that conferences are the main publication venues in information systems. The growth in interest in publishing in journals could indicate a shift in the depth of the research.

- EC16: Nearly similar numbers of papers were published in journals and conferences during 2021.

No significant trends can be detected from annual changes in the research focus. The research focusing on black boxes seemed to gain in popularity, whereas the research with the main focus on biases seemed to decrease in popularity. The number of papers focusing on attitudes seemed to grow relatively steadily. From the attitude papers, the annual division of papers focusing on understanding the developers and practitioners was as follows: one paper in 2018, two papers in 2019, six papers in 2020, and four papers in 2021. Understanding the expectations, needs, and opinions of practitioners seems to be a slowly growing trend. This could indicate that the research field is increasingly interested in practical implementation.

- EC17: There is a growing interest in practical implementation and understanding of the needs and expectations of users and practitioners.

Out of 43 papers published in journals, 18 focused on attitudes. This is a large proportion of attitude papers, reaching 45% ($n = 40$). Since the rigor in journal publications is higher than that of

conference papers [86], this indicates that although the field lacks a plurality of studies on humans' role and attitudes, the quality of this type of research is high.

- EC18: The studies on the role and expectations of users and practitioners represent high-quality research.

This reflection may be explained according to the type of data used in the research. User research usually requires more time-consuming research methods. Therefore, the originality and quality of the evidence are higher, which fits better with the publication criteria of journals. This can be compared to the black-box papers, where 26% (8 papers) were published in journals, and the bias papers, where 23% (15 papers) were published in journals.

5.5 Analysis of Connection to Real-world Problems

To understand whether the study field focuses on real-world problems, the papers were evaluated based on the use of real-world data versus synthetic data. As mentioned at the end of Section 4, only 7% of papers (10 papers) used synthetic data. In addition, most of the papers described the connected real-world challenges in the introduction and background sections. Overall, the research field is close to real-world problems.

- PEC5: XAI researchers are interested in real-world problems and applications, not only technical aspects of the topic.

If the field of XAI research had been studied independently without the association of AI ethics, then the connection to real-world problems may have been different.

5.6 Summary of Empirical Contributions

Next, we summarize the empirical contributions and primary empirical contributions of this article. The article's main theoretical contribution is to map the research area, which supports future research by framing and visualizing the existing research. The secondary contribution comprises the PECs derived from the maps. The PECs are supplemented with ECs. ECs that were highlighted from the text body in previous sections are listed below.

- EC1: Most of the research papers in the field of AI ethics do not use empirical evidence. Only 23% of the papers provide empirical evidence.
- EC2: Empirical research on AI ethics grew significantly in 2018, corresponding with trends in public discourse.
- EC3: The most popular paper type in the research facet is a proposal for solving algorithmic bias.
- EC4: Almost one-fourth of the papers in the whole sample contribute to the research field with a computational solution to solve algorithmic biases.
- EC5: Half of the papers interested in users' and practitioners' attitudes and perceptions related to XAI and AI ethics are contributing by modeling and structuring the research area.
- EC6: The most prevalent paper type is a computational tool proposing a solution to a problem with bias. Every fifth paper presents this type of research.
- EC7: The research field seems a bit monotonous and immature when considering the variety of topics, research methods used, and contributions of the papers.
- EC8: Out of the papers focusing on black box ($n = 31$), 84% had full pertinence on XAI.
- EC9: The research field of XAI seems to be solution-oriented, and the research corpus with empirical evidences focuses more on finding solutions than exploring challenges.

- EC10: There is a research gap in understanding who is responsible for the actions of non-transparent systems and how the responsibility is decided and communicated.
- EC11: XAI is a young but growing empirical research area in the field of AI ethics.
- EC12: The research interest in XAI compared with all published empirical papers on AI ethics was highest in 2019. Since then, the interest in XAI has grown yearly but not as rapidly as the empirical research on AI ethics in general has.
- EC13: The trend is toward more practical implications and less philosophical framing of the focus area.
- EC14: The research contribution and interest seems to be shifting from proposing general computational solutions to proposing more holistic design/framework-level solutions and tools for specific application issues.
- EC15: The most popular publication venue is AIES, with 28% of papers published in it.
- EC16: Fairly similar numbers of papers were published in journals and conferences during 2021.
- EC17: There is a growing interest in practical implementation and understanding the needs and expectations of users and practitioners.
- EC18: Studies on the role and expectations of users and practitioners represent high-quality research.

The primary empirical contributions are listed below. In previous sections, the primary empirical contributions were listed from the text body to bring them to the reader's attention and ensure easy accessibility when skimming the paper. Primary empirical contributions are written in a context-enriched manner to support the understanding of readers who are not familiar with the full paper.

- PEC1: XAI is a significant research focus on the study field of AI ethics. Of the empirical research papers published after 2012, 28% are related to XAI.
- PEC2: In the study field of Ethical XAI, the most common type of empirical research is studying a novel technique that can solve a computational challenge.
- PEC3: The human perspective toward XAI is not well known. There is no in-depth understanding of the practitioners' and users' expectations and attitudes toward XAI.
- PEC4: Industrial implementation of XAI is not yet profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of XAI.
- PEC5: XAI researchers are interested in real-world problems and applications, not only technical aspects of the topic.

Theoretical and practical implications of the primary empirical contributions are evaluated next.

6 DISCUSSION

This section lists the proposals for the theoretical and practical implications of the PECs, which were the SMS process outcomes. In theoretical implications, PECs are reflected against the existing research. The practical implications are proposals and ideas for how the conclusions could be implemented in practice. The limitations of the research are discussed at the end of the section.

6.1 Theoretical Implications

The main theoretical implication of this article is the mapping of the research area presented in Section 5. The key outcomes of the analysis of the mapping process are in this section mirrored existing research. PECs are mirrored to the existing research and evaluated if they contradict or correspond to the existing research or provide a novel perspective. As the focus of this article is to understand the research area's scope and depth, rather than the quality of the articles, the primary

Primary Empirical Finding	Theoretical Implications
PEC1: XAI is a significant research focus on the study field of AI ethics. Of the empirical research papers published after 2012, 28% are related to XAI.	**Corresponding.** The number of XAI papers implies the importance of the research field and the emerging nature of interest. Results are corresponding to the importance of transparency issues [22].
PEC2: In the study field of Ethical XAI, the most common type of empirical research is studying a novel technique that can solve a computational challenge.	**Novel.** The sparsity of validation study shows the freshness of the research field.
PEC3: The human perspective toward XAI is not well known. There is no in-depth understanding of the practitioners' and users' expectations and attitudes toward XAI.	**Corresponding.** The same challenge was noted in previous research of Adadi and Berrada [15]
PEC4: Industrial implementation of XAI is not yet profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of XAI.	**Corresponding.** Related to PEC3 and to observations of Vakkuri et al. [45] in the research area of AI ethics in general.
PEC5: XAI researchers are interested in real-world problems and applications, not only technical aspects of the topic.	**Contradicting.** Even though black box research is distancing from the real-world problems [36] XAI research is close to real-world problems.

Fig. 8. Theoretical implications.

empirical contributions are related to those factors. The summary of the results is presented in Figure 8.

A significant proportion of papers related to XAI in the empirical research of AI ethics (PEC1) corresponds to the research of Jobin et al. [89], who noted that the transparency is the most frequently highlighted principle in AI ethics. Besides, the result reflects the overall importance and interest of XAI. At the same time, it illustrates XAI's connection to real-world problems, as it is studied with empirical methods.

The interest in proposing novel computational solutions (PEC2) shows the freshness in the field without practical results to validate. The research area of AI ethics holds interest in finding technical solutions to ethical problems [34], which correlates with a broader perspective. To our knowledge, there is no previous research that has analyzed the type of research done in the field, so the relation to existing research may be shallow.

Previous research has shown that the human role and perspective are understudied subjects, both from the user's and practitioner's point of view [2, 45, 57, 58]. The same finding was evident in this SMS (PEC3). Concerning the lack of research on users' and practitioners' expectations, there was a more specific gap with the lack of research on XAI's industrial implementation (PEC4). Vakkuri et al. [171] pointed out an analogous dilemma with AI ethics. Their research is one of the few papers cited in this SMS that aims to understand the current state of the practical implementation of ethical principles.

Unlike black-box problems where the research field is distanced from real-world problems [139], XAI makes a strong contribution to addressing real-world problems (PEC5). The vast majority of the papers focusing on black boxes used real-world data in their research. In addition, in most of the papers, societal issues were highlighted in background sections or introductions.

This article has brought some novel perspectives to the research area, contributed to existing research, and contradicted some prior perspectives. It is important to remember that in SMS, the papers are not studied as profoundly as they are in SLR. To form a more in-depth conclusion, the research should be continued with SLR, which could provide new insights.

6.2 Practical Implications

Some of the PECs only had a clear practical contribution. Hence, they are not analyzed by their relevance to practitioners. The research field has a close connection to real-world problems (PEC5).

Primary Empirical Finding	Practical Implication
PEC2: In the study field of Ethical XAI, the most common type of empirical research is studying a novel technique that can solve a computational challenge.	There are several open-source solution proposals modifiable to fit the company needs. There is significant potential for collaboration between academia and industry
PEC3: The human perspective toward XAI is not well known. There is no in-depth understanding of the practitioners' and users' expectations and attitudes toward XAI.	The solutions proposed in the research papers might not have practical implementation potential due to the lack of understanding the practical needs.
PEC4: Industrial implementation of XAI is not yet profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of XAI.	It is required to understand if and how the XAI solutions are needed and understood by business decision-makers to enable practical implementation.
PEC5: XAI researchers are interested in real-world problems and applications, not only technical aspects of the topic.	The research serves the practitioners looking for a specific solution, as the research is done with real-world data. The research is contributing to the societal and regulatory discussion.

Fig. 9. Practical implications.

The research provides knowledge and perspective to regulators and communicators by contributing to the field and tying the research to societal issues. For practitioners looking for specific solutions, the research area offers open-source models tested with real-world data that practitioners can benchmark and modify to fit their needs (PEC2 and PEC5). There are many practical solutions and models built in academia; hence, the collaboration potential between academia and practitioners is significant (PEC2).

In contrast to the above points, since the research field is new and emerging, a shortage of practical implementation is recognizable (PEC3 and PEC4). There is no guarantee that the research area's solution proposals have the potential to serve practitioners and users and ever be implemented into practice (PEC3). The current practical implementation level of XAI solutions is unknown, as well as the expectations or interest of business decision-makers. If decision-makers do not understand the need for XAI, then the practical implementation of XAI in businesses is not likely to happen on a bigger scale (PEC4). The summary of results is presented in Figure 9.

In conclusion, the analysis of practical implementation revealed the potential of even closer collaboration between practitioners and academia. At the same time, the research gap when it comes to understanding the perspectives of practitioners, users, or business decision makers can harm the practical implementation of XAI solutions. Overall, more research is required to advance knowledge and further develop the field.

6.3 Limitations of the Research

A common bias that systematic reviews suffer from is that positive outcomes are more likely to be published compared to negative ones [20]. Especially in the corpus of empirical research, this may lead to a lack of validation studies and leave out solutions that were not working as expected. The inclusion of conference proceedings is one solution to avoid publication bias [20]. Thus, bias should be decreased in this article.

The framing of the research question posed limitations to this study. Since the focus of the article was to understand the research field of AI ethics and the role of XAI in the field, the mapping undertaken for the present article provided this specific viewpoint. However, this viewpoint has its challenges, because the definition excludes all research papers with a focus on AI's interpretability without a clearly visible relation to ethical concerns.

Due to the variety of vocabulary used in these research topics, there is also uncertainty as to how accurately the used search keywords reflect the underlying research area. As the keywords were limited to ethics and its synonyms—and AI and its synonyms—there is a chance that key papers have been missed. These papers may be relevant, yet if there was no mention of AI and ethics

in the abstract, the papers were not included. Thus, it is important to note that for the primary search, the keywords could have been expanded to include responsible AI-related concepts and principles such as transparency and accountability.

There could have also been a larger scope of technology-centered terminology included in the search, for example, computer ethics, but while we have observed that computer ethics and AI ethics are related fields, they are still distinct from one another. Computer ethics is a branch of applied ethics that focuses on the ethical issues related to computer technology. It encompasses a wide range of issues, including privacy, security, intellectual property, access to information, and the impact of computer technology on society [38]. AI ethics, however, is a more specific field that focuses on the ethical issues arising from the development and use of **artificial intelligence (AI)** systems [89]. While both fields are concerned with ethical issues related to technology, AI ethics is a narrower and more specialized field, focusing specifically on the unique ethical challenges presented by AI systems.

We chose to scope our research to a 10-year period. There are some potential limitations that may arise. By leaving out research done prior to 2012, the study may miss historical events and developments that have shaped the research field, which may lead to an incomplete understanding of the field's evolution. We acknowledge that our study may lack historical context, but our focus was on recent trends and developments. It is possible that a 10-year period may not be enough to fully cover all the topics and issues in the field, potentially leading to a narrow or incomplete analysis. Also, the findings may not be generalizable to the entire AI ethics field. For these reasons, we narrowed down the scope to the role of XAI in the AI ethics research field to gain a more detailed view on the state of this sub-field in question. Through doing this, we found that the oldest papers added to the final sample were from 2017. Hence, leaving out research done prior to 2012 is justified, as it is unlikely to significantly affect the final sample.

During the primary search, some limitations were faced with the databases. Each database was screened, starting from the oldest papers, to track its potential changes during the screening process. However, with the larger databases of Scopus, Web of Science, and ProQuest, the number of hits varied between searches. Because of these problems, there is a chance that not all relevant papers were included. However, since all three multi-disciplinary databases were included in the study, it reduces the possibility that any relevant paper was missed.

After the primary search, the sample size ($N = 4{,}411$) was larger than expected, which limited the amount of attention dedicated to each paper during the screening and inclusion process. In other SMS studies, the initial take-in from separate databases has been significantly lower, for example, 1,062 papers [168], 1,769 papers [128], and 2,081 papers [27]. Due to the large sample, the literature search and inclusion processes were conducted mostly by one reviewer per paper. Thus, there is a chance for human error and false classification during the screening process. To ensure better quality, if the reviewer felt uncertain about a paper, then the paper was tagged, and another reviewer provided a second opinion. The papers included after each inclusion phase were re-evaluated during the following phases. However, the papers excluded during the early inclusion were not further evaluated, increasing the possibility that a suitable paper would be missing from the final study because of manual labeling failure.

7 FUTURE RESEARCH

There is potential to continue the SMS with the collected dataset to gain a more in-depth understanding of the AI ethics research field. The literature search and inclusion process was performed with clear guidelines and disciplinary following a stringent search process, which enables future use of the research material [93]. One potential research direction is to use the collected dataset with empirical evidence ($n = 503$) to observe other emerging themes in the research field of AI

ethics, such as health, education, or regulation. In the future, the results could be extended by expanding the dataset via adding new keywords in the primary search. Closely connected terms such as transparency and responsibility would provide deeper insight on the ethical perspective.

The SMS revealed research gaps in the existing corpus. There is a need to study how humans perceive XAI and what they are expecting from XAI systems, or whether they even value them at all. That knowledge could guide the research area to search for solutions that are needed. Cross-disciplinary research between computer scientists and humanists could continue to provide exciting insights to the field, as already demonstrated in research on the human perspective in AI (e.g., Reference [114]).

There is a shortage of understanding regarding users' and practitioners' expectations, needs, and attitudes toward XAI, and there was no research on the managerial perspective of XAI identified. A more profound understanding of the current implementation level is needed to ensure that the research has value for practitioners and business decision-makers. The research area would benefit from a more advanced understanding of industrial implementation of and especially the managerial perspective on transparent systems in companies using AI solutions. Top managers are the final decision-makers, and they are accountable for their products' actions. Moreover, they are the gatekeepers of funding for development. To ensure the solutions proposed in papers are implemented in practice, it is necessary to understand what business decision-makers want and where they are ready to invest.

8 CONCLUSION

In this article, the SMS method was utilized to visualize how XAI is researched in the field of AI ethics. SMS was chosen to provide a broader perspective on AI ethics, elucidate the research area in a more profound way, and clarify the role of XAI in the research area. There is potential to continue the SMS compiled in this article to gain a more in-depth understanding of the AI ethics research field or other emerging topics in the research area.

The expected findings included mapping of the covered topic and analysis of when, how, and why the research was done to reveal potential research gaps. The research question was, "What is the role of XAI in the research field of AI ethics?" and the three following sub-questions were identified:

- R1 What has been empirically researched in the field of AI ethics?
- R2 What is the state of published research on XAI in the field of AI ethics in the past 10 years?
- R3 Where are the research gaps in the field?

The main interest behind this article was XAI's practical implications. Hence, the research was narrowed to empirical papers.

A quick analysis of the dataset of empirical research in AI ethics ($n = 503$) revealed that, overall, the AI ethics research is rather theoretical, as only 23% of manually included papers ($n = 2{,}192$) used empirical evidence. Empirical research grew significantly in 2018. Since 2018, the empirical research has kept on growing each year. Similarly, the research focus in XAI grew significantly in 2018 and has kept growing ever since. The interest in XAI is a significant area in AI ethics research with empirical evidence, as 28% of the papers ($n = 503$) contributed to issues related to XAI.

In terms of its current state, XAI is a growing research area that is close to real-world problems. Most of the papers were more concerned with the technical or design perspective of the problem compared to the practical challenges in implementation. This indicates that XAI is still mainly interpreted as an academic challenge. The field would benefit from a more robust understanding of the needs, expectations, and attitudes of users and practitioners. Future research is required to understand how XAI is perceived by business decision-makers. This could help to take research findings and solutions to practice.

A APPENDIX

Tables 4–7 present the list of primary studies with the classification category. First author and the year of publication is used as an identifier, and the papers are included in the reference list.

Table 4. Overview of Primary Studies Part 1

1st Author	Research	Contribution	Focus	Pertinence
Caliskan et al. (2017) [39]	Proposal	Model	Bias	Partial
Babu et al. (2018) [21]	Proposal	Tool	Bias	Partial
Calmon et al. (2018) [40]	Proposal	Tool	Bias	Partial
Dixon et al. (2018) [52]	Proposal	Tool	Bias	Full
Ehsan et al. (2018) [55]	Proposal	Tool	Black Box	Full
Flexer et al. (2018) [59]	Validation	Model	Bias	Full
Grgić-Hlača et al. (2018) a [70]	Proposal	Procedure	Bias	Full
Grgić-Hlača et al. (2018) b [69]e	Experience	Model	Attitudes (users)	Partial
Henderson et al. (2018) [78]	Philosophical	Model	Bias	Partial
Iyer et al. (2018) [87]	Proposal	Tool	Black Box	Full
Raff et al. (2018) [135]	Proposal	Tool	Bias	Full
Shank et al. (2018) [150]	Philosophical	Model	Attitudes (users)	Marginal
Srivastava et al. (2018) [161]	Proposal	Procedure	Bias	Full
Veale et al. (2018) [176]	Experience	Model	Attitudes (practitioners)	Full
Yang et al. (2018) [192]	Proposal	Tool	Bias	Full
Zhang et al. (2018) [196]	Proposal	Tool	Bias	Full
Zhou et al. (2018) [199]	Philosophical	Model	Attitudes (users)	Marginal
Abeywickrama et al. (2019) [1]	Proposal	Procedure	Accountability	Partial
Addis et al. (2019) [4]	Experience	Advice	Attitudes (practitioners)	Full
Aïvodji et al. (2019) [18]	Proposal	Tool	Black Box	Full
Ali et al. (2019) [8]	Proposal	Tool	Bias	Full
Amini et al. (2019) [12]	Proposal	Tool	Bias	Partial
Barn (2019) [25]	Philosophical	Model	Attitudes (users)	Marginal
Beutel et al. (2019) [28]	Proposal	Tool	Bias	Partial
Bremner et al. (2019) [33]	Proposal	Tool	Black Box	Partial
Brunk et al. (2019) [35]	Proposal	Model	Black Box	Full
Cardoso et al. (2019) [97]	Proposal	Tool	Bias	Full
Celis et al. (2019) [43]	Validation	Model	Bias	Partial
Coston et al. (2019) [47]	Proposal	Tool	Bias	Full
Crockett et al. (2019) [48]	Philosophical	Model	Attitudes (users)	Partial
Garg et al. (2019) [64]	Proposal	Tool	Bias	Partial
Goel et al. (2019) [67]	Proposal	Tool	Bias	Partial
Green et al. (2019) [68]	Philosophical	Advice	Attitudes (users)	Marginal
Heidari et al. (2019) [76]	Philosophical	Advice	Bias	Partial
Hind et al. (2019) [81]	Proposal	Procedure	Black Box	Full
Kim et al. (2019) [92]	Proposal	Tool	Black Box	Full
Lai et al. (2019) [98]	Philosophical	Model	Attitudes (users)	Partial
Lakkaraju et al. (2019) [100]	Proposal	Procedure	Black Box	Full
Lux et al. (2019) [108]	Proposal	Tool	Bias	Full
Mitchell et al. (2019) [116]	Proposal	Procedure	Bias	Full
Noriega-Campero et al. (2019) [119]	Proposal	Tool	Bias	Full
Radovanović et al. (2019) [133]	Proposal	Specific solution	Bias	Partial
Raji et al. (2019) [136]	Validation	Tool	Bias	Full
Rubel et al. (2019) [138]	Philosophical	Model	Accountability	Full

Table 5. Overview of Primary Studies Part 2

1st Author	Research	Contribution	Focus	Pertinence
Saxena et al. (2019) [142]	Philosophical	Advice	Attitudes (users)	Marginal
Sivill (2019) [158]	Philosophical	Advice	Bias	Partial
Srinivasan et al. (2019) [160]	Proposal	Tool	Bias	Partial
Teso et al. (2019) [164]	Proposal	Procedure	Black Box	Full
Ustun et al. (2019) [167]	Proposal	Tool	Bias	Partial
Vakkuri et al. (2019) [169]	Experience	Procedure	Attitudes (practitioners)	Marginal
Vanderelst et al. (2019) [174]	Philosophical	Advice	Attitudes (users)	Marginal
Vetrò et al. (2019) [178]	Philosophical	Advice	Bias	Partial
Wang et al. (2019) [182]	Experience	Model	Attitudes (practitioners)	Marginal
Webb et al. (2019) [183]	Philosophical	Model	Attitudes (users)	Full
Wolf et al. (2019) [188]	Proposal	Model	Black Box	Full
Wouters et al. (2019) [189]	Experience	Model	Attitudes (users)	Partial
Yilmaz et al. (2019) [193]	Proposal	Tool	Black Box	Full
Adams et al. (2020) [3]	Proposal	Tool	Black Box	Full
Alonso et al. (2020) [9]	Validation	Tool	Black Box	Full
Asatiani et al. (2020) [15]	Proposal	Model	Black Box	Full
Aysolmaz et al. (2020) [17]	Experience	Procedure	Bias	Full
Balachander et al. (2020) [22]	Proposal	Specific solution	Black Box	Full
Balasubramaniam et al. (2020) [23]	Experience	Model	Attitudes (practitioners)	Partial
Belavadi et al. (2020) [26]	Proposal	Tool	Bias	Partial
Bowyer et al. (2020) [30]	Validation	Specific solution	Bias	Partial
Brandão et al. (2020) [31]	Proposal	Procedure	Bias	Full
Chakraborty et al. (2020) [44]	Proposal	Tool	Bias	Full
Chen et al. (2020) [46]	Proposal	Tool	Bias	Full
Clavell et al. (2020) [63]	Experience	Tool	Bias	Full
Cortés et al. (2020) [49]	Proposal	Procedure	Bias	Full
Dexe et al. (2020) [51]	Validation	Procedure	Attitudes (practitioners)	Partial
Haffar et al. (2020) [72]	Proposal	Tool	Black Box	Full
He et al. (2020) [75]	Proposal	Tool	Bias	Full
Helberger et al. (2020) [77]	Philosophical	Model	Attitudes (users)	Partial
Hong et al. (2020) [84]	Philosophical	Model	Attitudes (users)	Partial
Jo et al. (2020) [88]	Experience	Procedure	Bias	Marginal
Karpati et al. (2020) [90]	Philosophical	Advice	Black Box	Full
Kouvela et al. (2020) [94]	Proposal	Specific solution	Black Box	Partial
Lakkaraju et al. (2020) [99]	Proposal	Procedure	Black Box	Full
Leavy et al. (2020) [102]	Proposal	Tool	Bias	Partial
Loi et al. (2020) [105]	Validation	Procedure	Accountability	Marginal
Lonjarret et al. (2020) [107]	Proposal	Tool	Black Box	Full
Madaio et al. (2020) [109]	Experience	Model	Attitudes (practitioners)	Marginal
McDonald et al. (2020) [112]	Philosophical	Advice	Attitudes (users)	Partial
Mitchell et al. (2020) [115]	Proposal	Procedure	Bias	Partial
Nirav et al. (2020) [6]	Philosophical	Procedure	Attitudes (users)	Marginal
Oppold et al. (2020) [121]	Proposal	Procedure	Bias	Partial
Orr et al. (2020) [122]	Experience	Model	Attitudes (practitioners)	Partial
Paraschakis et al. (2020) [126]	Proposal	Specific solution	Bias	Partial

Table 6. Overview of Primary Studies Part 3

1st Author	Research	Contribution	Focus	Pertinence
Park et al. (2020) [127]	Proposal	Specific solution	Bias	Full
Percy et al. (2020) [129]	Proposal	Specific solution	Bias	Partial
Radovanović et al. (2020) [134]	Proposal	Tool	Bias	Full
Schelenz et al. (2020) [143]	Proposal	Specific solution	Attitudes (users)	Marginal
Schelter et al. (2020) [144]	Proposal	Procedure	Bias	Partial
Seizov et al. (2020) [147]	Experience	Model	Attitudes (users)	Partial
Sendak et al. (2020) [148]	Philosophical	Model	Black Box	Full
Sharma et al. (2020) [151]	Proposal	Procedure	Black Box	Full
Sharma, Zhang et al. (2020) [152]	Proposal	Tool	Bias	Partial
Shi et al. (2020) [154]	Validation	Specific solution	Bias	Partial
Shulman et al. (2020) [156]	Proposal	Tool	Black Box	Full
Slack et al. (2020) [159]	Proposal	Procedure	Black Box	Full
Srivastava et al. (2020) [162]	Proposal	Specific solution	Attitudes (users)	Marginal
Sun et al. (2020)	Proposal	Model	Bias	Partial
Vakkuri et al. (2020) a [171]	Experience	Model	Attitudes (practitioners)	Partial
Vakkuri et al. (2020) b [170]	Experience	Model	Attitudes (practitioners)	Partial
van Berkel et al. (2020) [172]	Philosophical	Advice	Attitudes (users)	Partial
Wilson et al. (2020) [186]	Experience	Model	Black Box	Marginal
Zhang, W. et al. (2020) [197]	Proposal	Tool	Bias	Full
Zhang, X. et al. (2020) [198]	Experience	Procedure	Bias	Partial
Albach et al. (2021) [7]	Philosophical	Advice	Attitudes (users)	Partial
Bandi et al. (2021) [24]	Proposal	Specific solution	Attitudes (users)	Marginal
Camacho et al. (2021) [41]	Experience	Advice	Attitudes (practitioners)	Partial
Gencoglu (2021) [66]	Proposal	Specific solution	Bias	Partial
Henriksen et al. (2021) [79]	Experience	Model	Accountability	Partial
Huynh et al. (2021) [85]	Proposal	Tool	Black Box	Full
Jacqueline et al. (2021) [73]	Philosophical	Advice	Attitudes (users)	Marginal
Li et al. (2021) [104]	Proposal	Tool	Bias	Full
Loi et al. (2021) [106]	Validation	Model	Accountability	Marginal
Mariotti et al. (2021) [110]	Proposal	Procedure	Black Box	Full
Pandey et al. (2021) [125]	Validation	Specific solution	Bias	Full
Perrier (2021)	Proposal	Tool	Bias	Partial
Puiu et al. (2021) [132]	Validation	Model	Black Box	Partial
Richardson et al. (2021) [137]	Validation	Procedure	Attitudes (practitioners)	Full
Schmid et al. (2021) [145]	Proposal	Procedure	Attitudes (users)	Marginal
Serban et al. (2021) [149]	Validation	Model	Attitudes (practitioners)	Partial
Stumpf et al. (2021) [163]	Proposal	Procedure	Attitudes (users)	Partial
van Stijn et al. (2021) [173]	Experience	Procedure	Bias	Partial
Wang et al. (2021) [181]	Philosophical	Model	Attitudes (users)	Marginal
Wilson et al. (2021) [187]	Proposal	Specific solution	Bias	Full
Yaghini et al. (2021) [190]	Proposal	Procedure	Attitudes (users)	Full
Yoshikawa et al. (2021) [194]	Proposal	Specific solution	Bias	Partial
Yu et al. (2021) [195]	Proposal	Specific solution	Bias	Partial
Zicari et al. (2021) [200]	Proposal	Procedure	Accountability	Partial
Aïvodji et al. (2021) [19]	Proposal	Tool	Bias	Full
Blanes-Selva et al. (2021) [29]	Proposal	Specific solution	Black Box	Marginal

Table 7. Overview of Primary Studies Part 4

1st Author	Research	Contribution	Focus	Pertinence
Breeden et al. (2021) [32]	Proposal	Tool	Bias	Full
da Silva et al. (2021) [50]	Proposal	Tool	Bias	Partial
Franco et al. (2021) [61]	Proposal	Procedure	Black Box	Full
Hartmann et al. (2021) [74]	Experience	Model	Attitudes (practitioners)	Full
Köchling et al. (2021) [96]	Validation	Model	Bias	Partial
Ortega et al. (2021) [123]	Proposal	Specific solution	Black Box	Full
Tomalin et al. (2021) [165]	Proposal	Procedure	Bias	Partial

REFERENCES

[1] Dhaminda B. Abeywickrama, Corina Cirstea, and Sarvapali D. Ramchurn. 2019. Model checking human-agent collectives for responsible AI. In *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'19)*. IEEE, 1–8. DOI: https://doi.org/10.1109/RO-MAN46459.2019.8956429

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* PP (09 2018), 1–1. DOI: https://doi.org/10.1109/ACCESS.2018.2870052

[3] Janet Adams and Hani Hagras. 2020. A type-2 fuzzy logic approach to explainable AI for regulatory compliance, fair customer outcomes and market stability in the global financial sector. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'20)*. IEEE, 1–8. DOI: https://doi.org/10.1109/FUZZ48607.2020.9177542

[4] Chiara Addis and Maria Kutar. 2019. AI management an exploratory survey of the influence of GDPR and FAT principles. In *IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI'19)*. IEEE, 342–347. DOI: https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00102

[5] HLEG AI. 2019. High-level expert group on artificial intelligence. European Commission. Available at https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[6] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in norm-aware agents. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'20)*. International Foundation for Autonomous Agents and Multiagent Systems, 16–24.

[7] Michele Albach and James R. Wright. 2021. *The Role of Accuracy in Algorithmic Process Fairness across Multiple Domains*. Association for Computing Machinery, New York, NY, 29–49.

[8] Junaid Ali, Muhammad Bilal Zafar, Adish Singla, and Krishna P. Gummadi. 2019. Loss-aversively fair classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 211–218. DOI: https://doi.org/10.1145/3306618.3314266

[9] Jose Alonso, Javier Toja-Alamancos, and Alberto Bugarín. 2020. Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'20)*. IEEE, 1–8. DOI: https://doi.org/10.1109/FUZZ48607.2020.9177770

[10] Ethem Alpaydin. 2014. *Introduction to Machine Learning* (3rd ed.). MIT Press, Cambridge, MA.

[11] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. 2020. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* 20, 1 (2020), 1–9.

[12] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. Uncovering and mitigating algorithmic bias through learned latent structure. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 289–295. DOI: https://doi.org/10.1145/3306618.3314243

[13] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Framling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'19)*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58 (2020), 82–115.

[15] Aleksandre Asatiani, Pekka Malo, Per Rådberg Nagbøl, Esko Penttinen, Tapani Rinta-Kahila, and Antti Salovaara. 2020. Challenges of explaining the behavior of black-box AI systems. *MIS Quart. Exec.* 19 (2020), 259–278.

[16] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563 (11 2018). DOI: https://doi.org/10.1038/s41586-018-0637-6

[17] Banu Aysolmaz, Deniz Iren, and Nancy Dau. 2020. Preventing algorithmic bias in the development of algorithmic decision-making systems: A delphi study. In *53rd Hawaii International Conference on System Sciences. (HICSS'20)*. 5267–5276. DOI: https://doi.org/10.24251/HICSS.2020.648

[18] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: The risk of rationalization. In *36th International Conference on Machine Learning*. arxiv:1901.09749.

[19] Ulrich Aïvodji, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu, and Alain Tapp. 2021. Local data debiasing for fairness based on generative adversarial training. *Algorithms* 14, 3 (2021), 87. DOI: https://doi.org/10.3390/a14030087

[20] B. A. Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Citeseer. Retrieved from https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf.

[21] Mylam Babu and Sankaralingam Pushpa. 2018. An efficient discrimination prevention and rule protection algorithms avoid direct and indirect data discrimination in web mining. *Int. J. Intell. Eng. Syst.* 11 (08 2018), 212–220. DOI: https://doi.org/10.22266/ijies2018.0831.21

[22] T. Balachander, Aman Batra, and M. Choudhary. 2020. Machine learning pipeline for an improved medical decision support. *Int. J. Adv. Sci. Technol.* 29 (2020), 2632–2640. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/13723/6999.

[23] Nagadivya Balasubramaniam, Marjo Kauppinen, Sari Kujala, and Kari Hiekkanen. 2020. *Ethical Guidelines for Solving Ethical Issues and Developing AI Systems*. Springer, Cham, 331–346. DOI: https://doi.org/10.1007/978-3-030-64148-1_21

[24] Harit Bandi, Suyash Joshi, Siddhant Bhagat, and Dayanand Ambawade. 2021. Integrated technical and sentiment analysis tool for market index movement prediction, comprehensible using XAI. In *International Conference on Communication Information and Computing Technology (ICCICT'21)*. IEEE, 1–8. DOI: https://doi.org/10.1109/ICCICT50803.2021.9510124

[25] Balbir Barn. 2019. Mapping the public debate on ethical concerns: Algorithms in mainstream media. *J. Inf., Commun. Ethics Societ.* 18, 1 (2019), 124–139. DOI: https://doi.org/10.1108/JICES-04-2019-0039

[26] Vibha Belavadi, Yan Zhou, Murat Kantarcioglu, and Bhavani Thuriasingham. 2020. *Attacking Machine Learning Models for Social Good*. Springer, Cham, 457–471. DOI: https://doi.org/10.1007/978-3-030-64793-3_25

[27] L. Belmonte, R. Morales, and Antonio Fernández-Caballero. 2019. Computer vision in autonomous unmanned aerial vehicles—A systematic mapping study. *Appl. Sci.* 9 (08 2019), 3196. DOI: https://doi.org/10.3390/app9153196

[28] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 453–459. DOI: https://doi.org/10.1145/3306618.3314234

[29] Vicent Blanes-Selva, Ascensión Doñate-Martínez, Gordon Linklater, Jorge Garcés-Ferrer, and Juan M. García-Gómez. 2021. Responsive and minimalist app based on explainable AI to assess palliative care needs during bedside consultations on older patients. *Sustainability* 13, 17 (2021). DOI: https://doi.org/10.3390/su13179844

[30] Kevin W. Bowyer, Michael C. King, Walter J. Scheirer, and Kushal Vangara. 2020. The "Criminality from Face" illusion. *IEEE Trans. Technol. Societ.* 1, 4 (2020), 175–183. DOI: https://doi.org/10.1109/TTS.2020.3032321

[31] Martim Brandão, Marina Jirotka, Helena Webb, and Paul Luff. 2020. Fair navigation planning: A resource for characterizing and designing fairness in mobile robots. *Artif. Intell.* 282 (2020), 103259. DOI: https://doi.org/10.1016/j.artint.2020.103259

[32] Joseph L. Breeden and Eugenia Leonova. 2021. Creating unbiased machine learning models by design. *J. Risk Finan. Manag.* 14, 11 (2021). DOI: https://doi.org/10.3390/jrfm14110565

[33] Paul Bremner, Louise A. Dennis, Michael Fisher, and Alan F. Winfield. 2019. On proactive, transparent, and verifiable ethical reasoning for robots. *Proc. IEEE* 107, 3 (2019), 541–561. DOI: https://doi.org/10.1109/JPROC.2019.2898267

[34] Miles Brundage. 2014. Limitations and risks of machine ethics. *J. Experim. Theoret. Artif. Intell.* 26 (06 2014). DOI: https://doi.org/10.1080/0952813X.2014.895108

[35] Jens Brunk, Jana Mattern, and Dennis M. Riehle. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *IEEE 21st Conference on Business Informatics (CBI)*. IEEE, 429–435. DOI: https://doi.org/10.1109/CBI.2019.00056

[36] Joanna J. Bryson. 2019. The past decade and future of AI's impact on society. *Tow. New Enlight.* 11 (2019), 150–185.

[37] David Budgen, Mark Turner, Pearl Brereton, and Barbara Kitchenham. 2008. Using mapping studies in software engineering. *Proc. PPIG 2008* 2 (01 2008).

[38] Terrell Bynum. 2018. Computer and information ethics. In *The Stanford Encyclopedia of Philosophy (Spring 2018 ed.)*, Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[39] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. DOI: https://doi.org/10.1126/science.aal4230.

[40] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2018. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE J. Select. Topics Sig. Process.* 12, 5 (2018), 1106–1119. DOI: https://doi.org/10.1109/JSTSP.2018.2865887

[41] Javier Camacho and Mónica Olmeda. 2021. Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI Societ.* (08 2021). DOI: https://doi.org/10.1007/s00146-021-01267-0

[42] Cansu Canca. 2020. Operationalizing AI ethics principles. *Commun. ACM* 63, 12 (Nov. 2020), 18–21. DOI: https://doi.org/10.1145/3430368

[43] Diego Celis and Meghana Rao. 2019. Learning facial recognition biases through VAE latent representations. In *Conference on Fairness, Accountability, and Transparency, and Transparency in MultiMedia (FAT/MM'19)*. Association for Computing Machinery, New York, NY, 26–32. DOI: https://doi.org/10.1145/3347447.3356752

[44] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. *Fairway: A Way to Build Fair ML Software*. Association for Computing Machinery, New York, NY, 654–665.

[45] Jiahao Chen and Victor Storchan. 2021. Seven challenges for harmonizing explainability requirements. *arXiv preprint arXiv:2108.05390* (2021).

[46] Violet (Xinying) Chen and J. N. Hooker. 2020. *A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism*. Association for Computing Machinery, New York, NY, 221–227. DOI: https://doi.org/10.1145/3375627.3375844

[47] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 91–98. DOI: https://doi.org/10.1145/3306618.3314236

[48] Keeley Crockett, Sean Goltz, Matt Garratt, and Annabel Latham. 2019. Trust in computational intelligence systems: A case study in public perceptions. In *IEEE Congress on Evolutionary Computation (CEC'19)*. IEEE, 3227–3234. DOI: https://doi.org/10.1109/CEC.2019.8790147

[49] Efrén Cruz Cortés and Debashis Ghosh. 2020. *An Invitation to System-wide Algorithmic Fairness*. Association for Computing Machinery, New York, NY, 235–241.

[50] Daniela America da Silva, Henrique Duarte Borges Louro, Gildarcio Sousa Goncalves, Johnny Cardoso Marques, Luiz Alberto Vieira Dias, Adilson Marques da Cunha, and Paulo Marcelo Tasinaffo. 2021. Could a conversational AI identify offensive language? *Information* 12, 10 (2021). DOI: https://doi.org/10.3390/info12100418

[51] Jacob Dexe, Ulrik Franke, Anneli Avatare Nöu, and Alexander Rad. 2020. Towards increased transparency with value sensitive design. In *1st International Conference on Artificial Intelligence in HCI, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020*. Springer-Verlag, Berlin, 3–15. DOI: https://doi.org/10.1007/978-3-030-50334-5_1

[52] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. Association for Computing Machinery, New York, NY, 67–73. DOI: https://doi.org/10.1145/3278721.3278729

[53] Filip Karlo Došilovic, Mario Brcic, and Nikica Hlupic. 2018. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'18)*. 0210–0215.

[54] Filip Došilović, Mario Brcic, and Nikica Hlupic. 2018. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'18)*. IEEE, 0210–0215.

[55] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. Association for Computing Machinery, New York, NY, 81–87. DOI: https://doi.org/10.1145/3278721.3278736

[56] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I. Lee, Michael Muller, Mark O. Riedl, et al. 2021. The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

[57] Upol Ehsan and Mark O. Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[58] Juliana Ferreira and Mateus Monteiro. 2020. What are people doing about XAI user experience? A survey on AI explainability research and practice. In *International Conference on Human-Computer Interaction*. Springer, 56–73. DOI: https://doi.org/10.1007/978-3-030-49760-6_4

[59] Arthur Flexer, Monika Dörfler, Jan Schlüter, and Thomas Grill. 2018. Hubness as a case of technical algorithmic bias in music recommendation. In *IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1062–1069. DOI: https://doi.org/10.1109/ICDMW.2018.00154

[60] Luciano Floridi and Andrew Strait. 2020. *Ethical Foresight Analysis: What It Is and Why It Is Needed?* Vol. 3. Kluwer Academic Publishers, 77–97. DOI: https://doi.org/10.1007/s11023-020-09521-y

[61] Danilo Franco, Luca Oneto, Nicolò Navarin, and Davide Anguita. 2021. Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition. *Entropy* 23, 8 (2021). DOI: https://doi.org/10.3390/e23081047

[62] Jordan D. Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, and Maryellen L. Giger. 2021. A review of explainable and interpretable AI with applications in COVID-19 imaging. *Med. Phys.* 49, 1 (2021), 1–14.

[63] Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. 2020. *Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization.* Association for Computing Machinery, New York, NY, 265–271. DOI: https://doi.org/10.1145/3375627.3375852

[64] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19).* Association for Computing Machinery, New York, NY, 219–226. DOI: https://doi.org/10.1145/3306618.3317950

[65] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* 106 (2019), 101–121. DOI: https://doi.org/10.1016/j.infsof.2018.09.006

[66] Oguzhan Gencoglu. 2021. Cyberbullying detection with fairness constraints. *IEEE Internet Comput.* 25, 1 (2021), 20–29. DOI: https://doi.org/10.1109/MIC.2020.3032461

[67] Naman Goel and Boi Faltings. 2019. Crowdsourcing with fairness, diversity and budget constraints. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19).* Association for Computing Machinery, New York, NY, 297–304. DOI: https://doi.org/10.1145/3306618.3314282

[68] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-comput. Interact.* 3, CSCW (Nov. 2019). DOI: https://doi.org/10.1145/3359152

[69] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *World Wide Web Conference (WWW'18).* International World Wide Web Conferences Steering Committee, 903–912. DOI: https://doi.org/10.1145/3178876.3186138

[70] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI Conference on Artificial Intelligence.*

[71] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5 (2018), 1–42.

[72] Rami Haffar, Josep Domingo-Ferrer, and David Sánchez. 2020. *Explaining Misclassification and Attacks in Deep Learning via Random Forests.* Springer, Cham, 273–285. DOI: https://doi.org/10.1007/978-3-030-57524-3_23.

[73] Jacqueline Hannan, Huei-Yen Winnie Chen, and Kenneth Joseph. 2021. *Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation.* Association for Computing Machinery, New York, NY, 555–565. https://doi.org/10.1145/3461702.3462568

[74] Kathrin Hartmann and Georg Wenzelburger. 2021. Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Polic Sci.* 54 (06 2021). DOI: https://doi.org/10.1007/s11077-020-09414-y

[75] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. *A Geometric Solution to Fair Representations.* Association for Computing Machinery, New York, NY, 279–285. DOI: https://doi.org/10.1145/3375627.3375864

[76] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Conference on Fairness, Accountability, and Transparency (FAT*'19).* Association for Computing Machinery, New York, NY, 181–190. DOI: https://doi.org/10.1145/3287560.3287584

[77] Natali Helberger, Theo Araujo, and Claes H. de Vreese. 2020. Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Comput. Law Secur. Rev.* 39 (2020), 105456. DOI: https://doi.org/10.1016/j.clsr.2020.105456

[78] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18).* Association for Computing Machinery, New York, NY, 123–129. DOI: https://doi.org/10.1145/3278721.3278777

[79] Anne Henriksen, Simon Enni, and Anja Bechmann. 2021. *Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI.* Association for Computing Machinery, New York, NY, 574–585. DOI: https://doi.org/10.1145/3461702.3462564

[80] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossr. ACM Mag. Students* 25, 3 (2019), 16–19.

[81] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2019. TED: Teaching AI to explain its decisions. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 123–129. DOI: https://doi.org/10.1145/3306618.3314273

[82] Elizabeth A. Holm. 2019. In defense of the black box. *Science* 364, 6435 (2019), 26–27. DOI: https://doi.org/10.1126/science.aax0162

[83] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev.: Data Mining Knowl. Discov.* 9, 4 (2019), e1312.

[84] Joo-Wha Hong, Sukyoung Choi, and Dmitri Williams. 2020. Sexist AI: An experiment integrating CASA and ELM. *Int. J. Hum.-comput. Interact.* 36 (08 2020). DOI: https://doi.org/10.1080/10447318.2020.1801226

[85] Trung Dong Huynh, Niko Tsakalakis, Ayah Helal, Sophie Stalla-Bourdillon, and Luc Moreau. 2021. Addressing regulatory requirements on explanations for automated decisions with provenance—A case study. *Digit. Gov.: Res. Pract.* 2, 2 (Jan 2021). DOI: https://doi.org/10.1145/3436897

[86] Martin Ivarsson and Tony Gorschek. 2011. A method for evaluating rigor and industrial relevance of technology evaluations. *Empir. Softw. Eng.* 16 (06 2011), 365–395. DOI: https://doi.org/10.1007/s10664-010-9146-4

[87] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. 2018. Transparency and explanation in deep reinforcement learning neural networks. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. Association for Computing Machinery, New York, NY, 144–150. DOI: https://doi.org/10.1145/3278721.3278776

[88] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Conference on Fairness, Accountability, and Transparency (FAT*'20)*. Association for Computing Machinery, New York, NY, 306–316. DOI: https://doi.org/10.1145/3351095.3372829

[89] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1 (09 2019). DOI: https://doi.org/10.1038/s42256-019-0088-2

[90] Daniel Karpati, Amro Najjar, and Diego Agustin Ambrossio. 2020. *Ethics of Food Recommender Applications*. Association for Computing Machinery, New York, NY, 313–319.

[91] Emre Kazim and Adriano Soares Koshiyama. 2021. A high-level overview of AI ethics. *Patterns* 2, 9 (2021), 100314.

[92] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 247–254. DOI: https://doi.org/10.1145/3306618.3314287

[93] Barbara A. Kitchenham, David Budgen, and O. Pearl Brereton. 2011. Using mapping studies as the basis for further research—A participant-observer case study. *Inf. Softw. Technol.* 53, 6 (2011), 638–651. DOI: https://doi.org/10.1016/j.infsof.2010.12.011

[94] Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. 2020. Bot-detective: An explainable Twitter bot detection service with crowdsourcing functionalities. In *12th International Conference on Management of Digital EcoSystems (MEDES'20)*. Association for Computing Machinery, New York, NY, 55–63. DOI: https://doi.org/10.1145/3415958.3433075

[95] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* (2022).

[96] Alina Köchling, Shirin Riazy, Marius Claus Wehner, and Katharina Simbeck. 2021. Highly accurate, but still discriminatory. *Bus. Inf. Syst. Eng.: Int. J. Wirtschaftsinformatik* 63, 1 (2021), 39–54. DOI: https://doi.org/10.1007/s12599-020-00673-w

[97] Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. 2019. A framework for benchmarking discrimination-aware models in machine learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 437–444. DOI: https://doi.org/10.1145/3306618.3314262

[98] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Conference on Fairness, Accountability, and Transparency (FAT*'19)*. Association for Computing Machinery, New York, NY, 29–38. DOI: https://doi.org/10.1145/3287560.3287590

[99] Himabindu Lakkaraju and Osbert Bastani. 2020. *"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations*. Association for Computing Machinery, New York, NY, 79–85.

[100] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 131–138. DOI: https://doi.org/10.1145/3306618.3314229

[101] Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet Polic Rev.* 9, 2 (2020), 1–16.

[102] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. 2020. *Mitigating gender bias in machine learning data sets*. In *International Workshop on Algorithmic Bias in Search and Recommendation*. 12–26. DOI: https://doi.org/10.1007/978-3-030-52485-2_2

[103] Yann LeCun, Y. Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521 (05 2015), 436–44. DOI: https://doi.org/10.1038/nature14539

[104] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. *User-oriented Fairness in Recommendation.* Association for Computing Machinery, New York, NY, 624–632. DOI: https://doi.org/10.1145/3442381.3449866

[105] Michele Loi, Christoph Heitz, and Markus Christen. 2020. A comparative assessment and synthesis of twenty ethics codes on AI and big data. In *7th Swiss Conference on Data Science (SDS'20).* IEEE, 41–46. DOI: https://doi.org/10.1109/SDS49233.2020.00015

[106] Michele Loi and Matthias Spielkamp. 2021. *Towards Accountability in the Use of Artificial Intelligence for Public Administrations.* Association for Computing Machinery, New York, NY, 757–766.

[107] Corentin Lonjarret, Céline Robardet, Marc Plantevit, Roch Auburtin, and Martin Atzmueller. 2020. Why should I trust this item? Explaining the recommendations of any model. In *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA'20).* IEEE, 526–535. DOI: https://doi.org/10.1109/DSAA49011.2020.00067

[108] Thomas C. H. Lux, Stefan Nagy, Mohammed Almanaa, Sirui Yao, and Reid Bixler. 2019. A case study on a sustainable framework for ethically aware predictive modeling. In *IEEE International Symposium on Technology and Society (ISTAS'19).* IEEE, 1–7. DOI: https://doi.org/10.1109/ISTAS48451.2019.8937885

[109] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. *Co-designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI.* Association for Computing Machinery, New York, NY, 1–14.

[110] Ettore Mariotti, Jose M. Alonso, and Roberto Confalonieri. 2021. A framework for analyzing fairness, accountability, transparency and ethics: A use-case in banking services. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'21).* IEEE, 1–6. DOI: https://doi.org/10.1109/FUZZ45933.2021.9494481

[111] John McCarthy. 1959. Programs with common sense. In *Teddington Conference on the Mechanization of Thought Processes.* Her Majesty's Stationary Office, London, 75–91. Retrieved from http://www-formal.stanford.edu/jmc/mcc59.html.

[112] Nora McDonald and Shimei Pan. 2020. Intersectional AI: A study of how information science students think about ethics and their impact. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (Oct. 2020). DOI: https://doi.org/10.1145/3415218

[113] István Mezgár. 2021. From ethics to standards; an overview of AI ethics in CPPS. *IFAC-PapersOnLine* 54, 1 (2021), 723–728.

[114] Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (06 2017). DOI: https://doi.org/10.1016/j.artint.2018.07.007

[115] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. *Diversity and Inclusion Metrics in Subset Selection.* Association for Computing Machinery, New York, NY, 117–123.

[116] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Conference on Fairness, Accountability, and Transparency (FAT*'19).* Association for Computing Machinery, New York, NY, 220–229. DOI: https://doi.org/10.1145/3287560.3287596

[117] Tom M. Mitchell. 1980. *The Need for Biases in Learning Generalizations.* Technical Report. Rutgers, New Brunswick, NJ.

[118] J. Morley, L. Floridi, Libby Kinsey, and Anat Elhalal. 2019. From what to how. An overview of AI ethics tools, methods and research to translate principles into practices. *ArXiv* abs/1905.06876 (2019).

[119] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex "Sandy" Pentland. 2019. Active fairness in algorithmic decision making. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19).* Association for Computing Machinery, New York, NY, 77–83. DOI: https://doi.org/10.1145/3306618.3314277

[120] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,.* 1st Ed., IEEE. Available at https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf.

[121] Sarah Oppold and Melanie Herschel. 2020. A system framework for personalized and transparent data-driven decisions. In *32nd International Conference on Advanced Information Systems Engineering.* Springer-Verlag, Berlin, 153–168. DOI: https://doi.org/10.1007/978-3-030-49435-3_10

[122] Will Orr and Jenny L. Davis. 2020. Attributions of ethical responsibility by artificial intelligence practitioners. *Inf., Commun. Societ.* 23, 5 (2020), 719–735. DOI: https://doi.org/10.1080/1369118X.2020.1713842.

[123] Alfonso Ortega, Julian Fierrez, Aythami Morales, Zilong Wang, Marina Cruz, César Alonso, and Tony Ribeiro. 2021. Symbolic AI for XAI: Evaluating LFIT inductive programming for explaining biases in machine learning. *Computers* 10 (11 2021), 154. DOI: https://doi.org/10.3390/computers10110154

[124] Leila Ouchchy, Allen Coin, and Veljko Dubljevic. 2020. AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI Societ.* 35 (12 2020). DOI: https://doi.org/10.1007/s00146-020-00965-5

[125] Akshat Pandey and Aylin Caliskan. 2021. *Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms.* Association for Computing Machinery, New York, NY, 822–833. DOI: https://doi.org/10.1145/3461702.3462561

[126] Dimitris Paraschakis and Bengt Nilsson. 2020. *Matchmaking under Fairness Constraints: A Speed Dating Case Study.* Springer, 43–57. DOI: https://doi.org/10.1007/978-3-030-52485-2_5

[127] Sungho Park, Sunhee Hwang, Jongkwang Hong, and Hyeran Byun. 2020. Fair-VQA: Fairness-aware visual question answering through sensitive attribute prediction. *CEUR Worksh. Proc.* 8 (2020), 215091–215099. DOI: https://doi.org/10.1109/ACCESS.2020.3041503

[128] Nicolò Paternoster, Carmine Giardino, Michael Unterkalmsteiner, Tony Gorschek, and Pekka Abrahamsson. 2014. Software development in startup companies: A systematic mapping study. *Inf. Softw. Technol.* 56 (10 2014). DOI: https://doi.org/10.1016/j.infsof.2014.04.014

[129] Christian Percy, Artur d'Avila Garcez, Simo Dragicevic, and Sanjoy Sarkar. 2020. Lessons learned from problem gambling classification: Indirect discrimination and algorithmic fairness*. *Proceedings of the AAAI Fall Symposium on AI for Social Good*, Vol. 2884. CEUR Workshop Proceedings. Available at https://ceur-ws.org/Vol-2884/paper_107.pdf.

[130] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering.*

[131] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Inf. Softw. Technol.* 64 (2015), 1–18. DOI: https://doi.org/10.1016/j.infsof.2015.03.007

[132] Andrei Puiu, Anamaria Vizitu, Cosmin Nita, Lucian Mihai Itu, Puneet Sharma, and Dorin Comaniciu. 2021. Privacy-preserving and explainable AI for cardiovascular imaging. *Stud. Inform. Contr.* 30 (06 2021), 21–32. DOI: https://doi.org/10.24846/v30i2y202102

[133] Sandro Radovanović, Andrija Petrović, Boris Delibašić, and Milija Suknović. 2019. Making hospital readmission classifier fair—What is the cost? In *Central European Conference on Information and Intelligent Systems.*

[134] Sandro Radovanović, Andrija Petrović, Boris Delibašić, and Milija Suknović. 2020. Enforcing fairness in logistic regression algorithm. In *International Conference on INnovations in Intelligent SysTems and Applications (INISTA'20).* IEEE, 1–7. DOI: https://doi.org/10.1109/INISTA49547.2020.9194676

[135] Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair forests: Regularized tree induction to minimize model bias. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18).* Association for Computing Machinery, New York, NY, 243–250. DOI: https://doi.org/10.1145/3278721.3278742

[136] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19).* Association for Computing Machinery, New York, NY, 429–435. DOI: https://doi.org/10.1145/3306618.3314244

[137] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ML toolkits. In *CHI Conference on Human Factors in Computing Systems (CHI'21).* Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3411764.3445604

[138] Alan Rubel, Clinton Castro, and Adam Pham. 2019. Agency laundering and information technologies. *Ethic. Theor. Moral Pract.* 22 (08 2019), 1–25. DOI: https://doi.org/10.1007/s10677-019-10030-w

[139] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1 (05 2019), 206–215. DOI: https://doi.org/10.1038/s42256-019-0048-x

[140] Cynthia Rudin and Joanna Radin. 2019. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* 1, 2 (2019). https://doi.org/10.1162/99608f92.5a8a3a3d

[141] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU J.: ICT Discov. - Special Iss. 1 - Impact Artif. Intell. (AI) Commun. Netw. Serv.* 1 (10 2017), 1–10.

[142] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19).* Association for Computing Machinery, New York, NY, 99–106. DOI: https://doi.org/10.1145/3306618.3314248

[143] Laura Schelenz, Avi Segal, and Kobi Gal. 2020. Best practices for transparency in machine generated personalization. In *28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP'20 Adjunct).* Association for Computing Machinery, New York, NY, 23–28. DOI: https://doi.org/10.1145/3386392.3397593

[144] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2020. FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Advances in Database Technology - EDBT 2020*, Angela Bonifati, Yongluan Zhou, Marcos Antonio Vaz Salles, Alexander Bohm, Dan Olteanu, George Fletcher, Arijit Khan, and Bin Yang (Eds.). OpenProceedings.org, 395–398. DOI: https://doi.org/10.5441/002/edbt.2020.41

[145] Thomas Schmid, Wolfgang Hildesheim, Taras Holoyad, and Kinga Schumacher. 2021. The AI methods, capabilities and criticality grid. *KI - Künstliche Intelligenz* 35 (2021), 425–440. DOI: https://doi.org/10.1007/s13218-021-00736-4

[146] Johannes Schneider, Joshua Handali, and Peter. 2019. Personalized explanation for machine learning: A conceptualization. *arXiv e-prints* (05 2019), arXiv:1901.00770.

[147] Ognyan Seizov and Alexander Wulf. 2020. Artificial intelligence and transparency: A blueprint for improving the regulation of AI applications in the EU. *Eur. Bus. Law Rev.* 31 (09 2020), 611–640.

[148] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": Supporting clinical decision-making with deep learning. In *Conference on Fairness, Accountability, and Transparency (FAT*'20)*. Association for Computing Machinery, New York, NY, 99–109. DOI: https://doi.org/10.1145/3351095.3372827

[149] Alex Serban, Koen van der Blom, Holger Hoos, and Joost Visser. 2021. Practices for engineering trustworthy machine learning applications. In *IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN'21)*. IEEE, 97–100. DOI: https://doi.org/10.1109/WAIN52551.2021.00021

[150] Daniel B. Shank and Alyssa DeSanti. 2018. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Comput. Hum. Behav.* 86 (2018), 401–411. DOI: https://doi.org/10.1016/j.chb.2018.05.014

[151] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. *CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models*. Association for Computing Machinery, New York, NY, 166–172.

[152] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. *Data Augmentation for Discrimination Prevention and Bias Disambiguation*. Association for Computing Machinery, New York, NY, 358–364.

[153] Mary Shaw. 2003. Writing good software engineering research paper, In *International Conference on Software Engineering*. 726–737.

[154] Sheng Shi, Shanshan Wei, Zhongchao Shi, Yangzhou Du, Wei Fan, Jianping Fan, Yolanda Conyers, and Feiyu Xu. 2020. *Algorithm Bias Detection and Mitigation in Lenovo Face Recognition Engine*. Springer International Publishing, 442–453. DOI: https://doi.org/10.1007/978-3-030-60457-8_36

[155] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.* 146 (2021), 102551.

[156] Eyal Shulman and Lior Wolf. 2020. *Meta Decision Trees for Explainable Recommendation Systems*. Association for Computing Machinery, New York, NY, 365–371.

[157] Moninder Singh, Gevorg Ghalachyan, Kush R. Varshney, and Reginald E. Bryant. 2021. An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness. *arXiv preprint arXiv:2109.14653* (2021).

[158] Torty Sivill. 2019. Ethical and statistical considerations in models of moral judgments. *Front. Robot. AI* 6 (2019). DOI: https://doi.org/10.3389/frobt.2019.00039

[159] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. *Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods*. Association for Computing Machinery, New York, NY, 180–186.

[160] Ramya Srinivasan and Ajay Chander. 2019. Understanding bias in datasets using topological data analysis. In *Workshop on Artificial Intelligence Safety, held at the International Joint Conference on Artificial Intelligence (AISafety@IJCAI'19)*. Retrieved from http://ceur-ws.org/Vol-2419/paper_9.pdf.

[161] Biplav Srivastava and Francesca Rossi. 2018. Towards composable bias rating of AI services. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. Association for Computing Machinery, New York, NY, 284–289. DOI: https://doi.org/10.1145/3278721.3278744

[162] Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi. 2020. Personalized chatbot trustworthiness ratings. *IEEE Trans. Technol. Societ.* 1, 4 (2020), 184–192. DOI: https://doi.org/10.1109/TTS.2020.3023919

[163] Simone Stumpf, Lorenzo Strappelli, Subeida Ahmed, Yuri Nakao, Aisha Naseer, Giulia Del Gamba, and Daniele Regoli. 2021. Design methods for artificial intelligence fairness and transparency. In *IUI Workshops*.

[164] Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 239–245. DOI: https://doi.org/10.1145/3306618.3314293

[165] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics Inf. Technol.* 23 (09 2021). DOI: https://doi.org/10.1007/s10676-021-09583-1

[166] Matteo Turilli and Luciano Floridi. 2009. The ethics of information transparency. *Ethics Inf. Technol.* 11 (06 2009), 105–112. DOI: https://doi.org/10.1007/s10676-009-9187-9

[167] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 6373–6382. Retrieved from https://proceedings.mlr.press/v97/ustun19a.html.

[168] Ville Vakkuri and Pekka Abrahamsson. 2018. The key concepts of ethics of artificial intelligence. In *IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC'18)*. IEEE, 1–6.

[169] Ville Vakkuri, Kai-Kristian Kemell, and P. Abrahamsson. 2019. Ethically aligned design: An empirical evaluation of the RESOLVEDD-strategy in software and systems development context. In *45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA'19)*. 46–50.

[170] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. 2020. *"This Is Just a Prototype": How Ethics Are Ignored in Software Startup-like Environments*. Springer International Publishing, Cham, 195–210. DOI: https://doi.org/10.1007/978-3-030-49392-9_13

[171] Ville Vakkuri, Kai-Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson. 2020. The current state of industrial practice in artificial intelligence ethics. *IEEE Softw.* PP (04 2020). DOI: https://doi.org/10.1109/MS.2020.2985621

[172] Niels van Berkel, Benjamin Tag, Jorge Goncalves, and Simo Hosio. 2020. Human-centred artificial intelligence: A contextual morality perspective. *Behav. Inf. Technol.* 41, 3 (2020), 502–518. DOI: https://doi.org/10.1080/0144929X.2020.1818828

[173] Jip J. van Stijn, Mark A. Neerincx, Annette ten Teije, and Steven Vethman. 2021. Team design patterns for moral decisions in hybrid intelligent systems: A case study of bias mitigation. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. CEUR-WS.

[174] Dieter Vanderelst and Jurgen Willems. 2020. Can we agree on what robots should be allowed to do? An exercise in rule selection for ethical care robots. *Int. J. Soc. Robot.* 12 (11 2020), 1093–1102. DOI: https://doi.org/10.1007/s12369-019-00612-0

[175] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Societ.* 4 (12 2017), 205395171774353. DOI: https://doi.org/10.1177/2053951717743530

[176] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. *Fairness and Accountability Design Needs for Algorithmic Support in High-stakes Public Sector Decision-making*. Association for Computing Machinery, New York, NY, 1–14. DOI: https://doi.org/10.1145/3173574.3174014

[177] Tom Vermeire, Thibault Laugel, Xavier Renard, David Martens, and Marcin Detyniecki. 2021. How to choose an explainability method? Towards a methodical implementation of XAI in practice. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 521–533.

[178] Antonio Vetro, Antonio Santangelo, Elena Beretta, and Juan Carlos De Martin. 2019. AI: From rational agents to socially responsible agents. *Digit. Polic., Regul. Govern.* 21 (02 2019). DOI: https://doi.org/10.1108/DPRG-08-2018-0049

[179] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. *arXiv preprint arXiv:2006.00093* (2020).

[180] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. J. Law Technol.* 31 (04 2018), 841–887. DOI: https://doi.org/10.2139/ssrn.3063289

[181] Weisha Wang, Long Chen, Mengran Xiong, and Yichuan Wang. 2021. Accelerating AI adoption with responsible AI signals and employee engagement mechanisms in health care. *Inf. Syst. Front.* (06 2021). DOI: https://doi.org/10.1007/s10796-021-10154-4

[182] Yichuan Wang, Mengran Xiong, and Hossein Olya. 2019. Toward an understanding of responsible artificial intelligence practices. DOI: https://doi.org/10.24251/HICSS.2020.610

[183] Helena Webb, Menisha Patel, Michael Rovatsos, Alan Davoust, Sofia Ceppi, Ansgar Koene, Liz Dowthwaite, Virginia Portillo, Marina Jirotka, and Monica Cano. 2019. "It would be pretty immoral to choose a random algorithm." Opening up algorithmic interpretability and transparency. *J. Inf., Commun. Ethics Societ.* 17, 2 (2019), 210–228. Retrieved from https://www.proquest.com/scholarly-journals/would-be-pretty-immoral-choose-random-algorithm/docview/2283793374/se-2?accountid=11774.

[184] Lindsay Wells and Tomasz Bednarz. 2021. Explainable AI and reinforcement learning—A systematic review of current approaches and trends. *Front. Artif. Intell.* 4 (2021), 48.

[185] Roel Wieringa, Neil Maiden, Nancy Mead, and Colette Rolland. 2006. Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. *Requir. Eng.* 11 (03 2006), 102–107. DOI: https://doi.org/10.1007/s00766-005-0021-6

[186] Campbell Wilson, Janis Dalins, and Gregory Rolan. 2020. Effective, explainable and ethical: AI for law enforcement and community safety. In *IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G'20)*. 186–191. DOI: https://doi.org/10.1109/AI4G50087.2020.9311021

[187] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening (FAccT'21). Association for Computing Machinery, New York, NY, 666–677. DOI: https://doi.org/10.1145/3442188.3445928

[188] Lior Wolf, Tomer Galanti, and Tamir Hazan. 2019. A formal approach to explainability. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. Association for Computing Machinery, New York, NY, 255–261. DOI: https://doi.org/10.1145/3306618.3314260

[189] Niels Wouters, Ryan Kelly, Eduardo Velloso, Katrin Wolf, Hasan Shahid Ferdous, Joshua Newn, Zaher Joukhadar, and Frank Vetere. 2019. Biometric mirror: Exploring ethical opinions towards facial analysis and automated decision-making. In *Designing Interactive Systems Conference (DIS'19)*. Association for Computing Machinery, New York, NY, 447–461. DOI: https://doi.org/10.1145/3322276.3322304

[190] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. 2021. *A Human-in-the-loop Framework to Construct Context-aware Mathematical Notions of Outcome Fairness*. Association for Computing Machinery, New York, NY, 1023–1033.

[191] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* 77 (2022), 29–52.

[192] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *International Conference on Management of Data (SIGMOD'18)*. Association for Computing Machinery, New York, NY, 1773–1776. DOI: https://doi.org/10.1145/3183713.3193568

[193] Levent Yilmaz and Sunit Sivaraj. 2019. A cognitive architecture for verifiable system ethics via explainable autonomy. In *IEEE International Systems Conference (SysCon'19)*. IEEE, 1–8. DOI: https://doi.org/10.1109/SYSCON.2019.8836896

[194] Hiroki Yoshikawa, Akira Uchiyama, and Teruo Higashino. 2021. Time-series physiological data balancing for regression. In *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA'21)*. IEEE, 393–398. DOI: https://doi.org/10.1109/ICAICA52286.2021.9498128

[195] Renzhe Yu, Hansol Lee, and René F. Kizilcec. 2021. Should college dropout prediction models include protected attributes? In *8th ACM Conference on Learning @ Scale (L@S'21)*. Association for Computing Machinery, New York, NY, 91–100. DOI: https://doi.org/10.1145/3430895.3460139

[196] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES'18)*. Association for Computing Machinery, New York, NY, 335–340. DOI: https://doi.org/10.1145/3278721.3278779

[197] Wenbin Zhang, Mingli Zhang, Ji Zhang, Zhen Liu, Zhiyuan Chen, Jianwu Wang, Edward Raff, and Enza Messina. 2020. Flexible and adaptive fairness-aware learning in non-stationary data streams. In *IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI'20)*. IEEE, 399–406. DOI: https://doi.org/10.1109/ICTAI50040.2020.00069

[198] Xi Zhang, Yuqing Zhao, Xinlin Tang, Hengshu Zhu, and Hui Xiong. 2020. Developing fairness rules for talent intelligence management system. In *Hawaii International Conference on System Sciences*. DOI: https://doi.org/10.24251/HICSS.2020.720

[199] Jianlong Zhou and Fang Chen. 2017. DecisionMind: Revealing human cognition states in data analytics-driven decision making with a multimodal interface. *J. Multimod. User Interf.* (10 2017). DOI: https://doi.org/10.1007/s12193-017-0249-8

[200] Roberto V. Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, Naveed Mushtaq, Gemma Roig, Norman Stürtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund. 2021. Z-Inspection®: A process to assess trustworthy AI. *IEEE Trans. Technol. Societ.* 2, 2 (2021), 83–97. DOI: https://doi.org/10.1109/TTS.2021.3066209

Received 11 February 2022; revised 24 February 2023; accepted 11 May 2023