Heidi Vainio-Pekka

THE ROLE OF EXPLAINABLE AI IN THE RESEARCH FIELD OF AI ETHICS – SYSTEMATIC MAPPING STUDY

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

ABSTRACT

Vainio-Pekka, Heidi

Master's Thesis
Jyväskylä: Jyväskylän yliopisto, 2020, 67 s.
Information System Science, Master Theses
Supervisors: Vakkuri Ville, Abrahamsson Pekka

This paper presents the Systemic Mapping Study results of the Ethics of Artificial Intelligence (AI) research. AI ethics is an emerging and versatile topic interesting to different domains. This paper focuses on understanding the role of Explainable AI in the research field and how the topic has been studied.

Explainable AI refers to AI systems that are interpretable or understandable to humans. It aims to increase the transparency of systems and make systems more trustworthy. Non-transparent AI systems are have already shown some of their weaknesses, such as in some cases favoring men over women in the hiring process.

The research fields of AI ethics and Explainable AI lack a common framework and conceptualization. There is no clarity of the field's depth and versatility; hence a systemic approach to understanding the corpus was needed. The systemic review offers an opportunity to detect research gaps and focus points. A Systemic Mapping Study is a tool to performing a repeatable and continuable literature search.

This paper contributes to the research field with a Systemic Map that visualizes what, how, when, and why Explainable AI has been studied in AI ethics. Within the scope is the detection of primary papers in AI ethics, which opens possibilities to continue the mapping process in other papers. The third contribution is the primary empirical conclusions drawn from the analysis and reflect existing research and practical implementation.

Keywords: AI Ethics, Explainable AI, Artificial Intelligence, Systemic Mapping Study

FIGURES

TABLES

CONTENT TABLE

1 INTRODUCTION

Artificial Intelligence (AI) is one of the most prominent and influential technologies of modern days. The importance of AI-empowered applications is predicted to grow in the future. Already today, AI is affecting the everyday life of common people from social media feed modifications and shopping recommendations to manipulation of people's voting preferences. The speed of development and the race between nations and companies to build robust AI tools increases the need to set the ethical guidelines and principles for AI development and deployment.

AI ethics is based on computer ethics, which is interested in human and machine interaction, and machine ethics, which is interested in moral agents and how morality can be programmed to the machines. AI ethics is often broken down to principles from which five of the most frequently required are transparency, justice, and fairness, non-maleficence, responsibility, and privacy (Jobin, Ienca and Vayena, 2019). Transparency, per se, can be seen as a pro-ethical principle, the enabler of ethical AI (Turilli and Floridi 2009). Explainable AI (later XAI) is aiming to solve the issues with transparency. XAI refers to the interpretable system that provides an understandable explanation to the system output (Adadi & Berrada, 2018). This paper aims to understand the research field of XAI and its role in AI ethics research.

1.1 Motivation

The subject of AI Ethics is versatile, ranging from the worries about conscious machines and their capability to replace people with machine workers, to more technical challenges such as designing ethical autonomous vehicles or settling the requirements of developing explainable machine learning algorithms. The field is broad and research areas vary from highly technical issues to understanding human behavior; hence it is a relevant research topic for social scientists, philosophers, economists, information system scientists, data scientists, mathematicians, and researchers from other domains. Multidisciplinary research is required to understand the research field's depth and extend and reveal potential research gaps.

Due to the novelty of the research area, it still lacks clarity and structure. AI ethics and XAI are both suffering from the lack of commonly agreed definitions of core concepts (Došilović, Brčić & Hlupić, 2018; Jobin et al., 2019). This paper aims to understand how XAI is researched from the perspective of AI ethics. This perspective requires first the understanding of the research field of AI ethics.

AI ethics is not just a future concern but a relevant issue of the real-world. Unfair non-transparent algorithmics are already in use (O'Neil, 2016). Mistakes by such algorithms may have long and unexpected consequences such as denials of university access (Evgeniou, Hardoon and Ovchinnikov, 2020). The issues are not just technical challenges, but a broader perspective is required. It is essential to understand the connection between real-world problems and academic research.

To understand what is researched in AI ethics and how XAI is presented in the research field, a study on the research corpus of AI ethics is required. This paper uses Systemic Mapping Study (later SMS) to map the research literature of AI ethics. The research question of this paper is: *What is the role of XAI in the research field of AI ethics?* divided into sub-questions:

[R1] What is researched in the AI ethics research field with empiric evidence?

[R2] What is the current state of XAI in the research field of AI ethics?

[R3] What are the research gaps in the field?

The sub-questions are opened and motivated in the following chapter, and the research method of SMS is shortly introduced next.

1.2 Research questions

This paper's research question *"What is the role of explainable AI in AI ethics' research field?"* requires an overview of the overall corpus of academic literature on AI ethics. As this paper is more focused on concrete issues rather than philosophical discussion, the focus is on the research with empirical evidence. To answer the research question, it is required first to answer the research question of *[R1] What is researched in the AI ethics research field with empiric evidence?* To profoundly answer this question, more in-depth research is required than what is possible to perform in a master's thesis. In this paper, the question is studied at a superficial level to offer enough background to understand the main research question. The major topics are noted, the research field's size, and the proportion of empiric research from the existing academic literature Further study is required to fully understand the full empiric research corpus of academic literature of AI ethics.

The second question is *[R2] What is the current state of XAI in the research field of AI ethics?* The research with XAI's focus is mirrored to a full dataset of empiric studies to understand XAI's role and importance in AI ethics. More profound analysis and classification are performed to papers focusing on XAI to understand when, what, how, and why it has been studied. 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.

The third dimension of the research question is to understand what has not been researched. The question *[R3] What are the research gaps in the field?* aims to answer that question based on background literature review and a profound SMS. Background literature review brought out gaps, such as the lack of understanding of the role of humans in XAI (Adadi and Berrada, 2018) that were also highlighted in SMS analysis. Other gaps were revealed, such as a lack of research of implementation in practice and the current state of XAI in organizations.

1.3 Research method

The research method applied in this work is the Systematic Mapping Study, SMS. The method is shortly introduced here and more profoundly explained simultaneously with the reporting of SMS used in the AI ethics research area. That allows the reader to follow SMS's theoretical framework and mirror it to this paper's application. Several SMS studied, and guidelines are utilized. However, the major contributing papers for this study are the guidelines of Petersen, Feldt, Mujtaba, and Mattsson (2008), and the SMS of Paternoster, Giardino, Unterkalmsteiner, Gorschek, and Abrahamsson (2014). This paper continues the SMS of Vakkuri and Abrahamsson (2018).

SMS is a form of Systematic Literature Review (SLR), which is a more commonly used literature review method. SLR and SMS are secondary studies where the attention is 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. Broader SMS aims to identify and categorize the existing literature. (Kitchenham, Budgen, and Brereton, 2011).

Standardly SLR has a specific, well-defined research question that can be answered with empiric research, wherein SMS typically has a broader view of the research topic. Another essential difference between SMS and SLR is that SLR has a stronger emphasis on the research outcomes of primary papers and analyzes their consistency. Wherein an SMS typically aims only to classify and categorize the relevant literature, and only the classification data is collected. The expected result of SMS is "a set of papers related to a topic area categorized in a variety of dimensions and counts of the number of papers in various categories." (Kitchenham et al., 2011).

To understand the role of XAI in the research field of AI ethics, SMS methodology served better than SLR. The freshness and incoherence of the AI Ethics research area advocated the use of SMS. The size of the research area was unknown, and the role of XAI new. Conceptual ambiguity of the research area (Jobin et al., 2019) supported SMS usage.

The benefit of SMS is the possibility of continuing the study to more in-depth SLR. That, though, requires the SMS to be a recent or updated well-reported high-quality work. To guarantee the quality, SMS must follow a stringent search process, snowball the primary study references, and have a well-defined calcification schema and process. SMS needs to be updated if it is not continued shortly after. The updating needs to follow the same procedure used in the original SMS. High-quality SMS can have a significant benefit for the research area in establishing the baselines for future research. (Kitchenham et al., 2011).

SMS is a highly time-consuming and rather challenging research method; hence it is not typically used in master's theses (Petersen et al., 2018). Undergraduates tend to lack the skills and academic understanding to produce high-quality SMS with future study opportunities. To guarantee the paper's quality, the topic must be carefully chosen to ensure a manageable number of included papers (Petersen et al., 2018). This paper was done in close collaboration and supervision of University of Jyväskylä's (JYU) AI Ethics Labs' research group to ensure the academic quality and validity. The literature search was performed with Vakkuri and Abrahamsson (2018) framework that ensured the quality of material gathering. The research area was significantly larger than expected, which challenged the rigor of the work. Part of the literature search and inclusion process was performed by a research assistant to keep the work-load manageable without jeopardizing the work's rigor and quality.

1.4 Structure of work

The first part of the paper serves as a background for the SMS. It presents the technologies AI and Machine Learning. Next, the ethical foundations and the principals for ethical AI are introduced. Following the introduction of XAI and related issues such as transparency and black box problem are described. At the end of the chapter, there is a short conclusion of the background study and the research area's analyses. This background chapter aims to provide the reader with an understanding of the topic of AI ethics and how the research area is interpreted.

The second part of the study reports the literature search process. The chapter starts with a theoretical framework of SMS and continues with reporting the use of SMS in this paper. The literature search is performed only for the year range of 2018-2020 to update the SMS by Vakkuri and Abrahamson (2018). After the literature search, the sample of 2018-2020 was 1975 papers.

In chapter three, the inclusion and exclusion criteria and the process is reported. The inclusion process was done during four screening rounds of the papers. After the first screening round, the sample of 2018-2020 (n=1532) was combined with a separately screened sample of 2012-2017 (n=403). After four screening rounds, the final dataset was 76 papers.

The identified primary studies (n=76) were analyzed during the next two chapters. Chapter four presents the classification schema and the numeric results of classification. In chapter five, the results are analyzed and compared, and the annual trends and the venues of publications are investigated. Chapter six is the Discussion where is proposed theoretical and practical implications of primary empirical conclusions. In Conclusions, the results are mirrored to the research questions, and the limitations of this study are analyzed—finally, future research topics are suggested.

2 BACKGROUND

The purpose of this background chapter is to present the main themes related to this study and highlight the current discussion around XAI. In addition, the aim is to provide the needed background knowledge for the reader. First, AI and related technologies are presented and followed by the ethical foundations and principles of AI. Next, XAI and related themes are shortly described. The chapter ends with conclusions and the motivation to proceed with the SMS.

2.1 Artificial Intelligence

The unambiguous definition of Artificial Intelligence is challenging. AI is used as an umbrella term for many technologies such as machine learning, machine vision, and autonomous machines. On the other hand, AI can be seen as part of the broader framework of digitalization. In academia, AI is a cross-disciplinary research area of engineering, economics, and humanistic sciences. In short, AI could be defined as, a tool that enables machines, programs, systems, and services to function rationally according to the task and situation (Russell & Norvig, 1994).

European Commission has taken an initiative to frame and regulate the use of AI. Their High-Level Expert Group on Artificial Intelligence, AI HLEG group, (Rossi et al., 2019) defines AI as follows:

> “Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and ro-

> botics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

Technology has been one of the fundamentals of economic growth ever since the industrial revolution. In the center of technological innovations are general-purpose technologies, such as a steam engine or electricity, that have the power to catalyze other complementary innovations. With the capability to improve itself without human intervention, AI is a general-purpose technology, making it a fascinating study subject. (Brynjolfsson & McAfee, 2017).

AI has a long history and has roots in the 60s, so it is far from being a new technology. During its history, AI has had its ups and downs in the hype curve, making it appear brand-new in public discussion. Despite the lack of hype in the industrial sector, AI has been a standard part of the industrial repertoire ever since the 80s (Bryson, 2019). However, it was not until 2007 that the introduction and generalization of smartphones and social media channels started to generate large amounts of data, which affected machine learning by providing it the training material and target applications (Bryson, 2019). Together with easier mass data access, the progress in computing power, and the development of Machine Learning algorithms the so-called Second Machine Age started. That brought AI back to the media spotlight and the top of the hype curve.

In general, the AI field suffers from overly high expectations regarding the speed of development and over-promised AI applications' capabilities. Even though technological development and increase in computing speed are constantly progressing, more time is required to get prominent AI systems from research laboratories to deployment-ready applications. Thus, too high expectations can cause disappointments and decrease interest in investments. The media and entertainment industries are filled with images of generally intelligent machines. However, generally intelligent AI is far from today's narrow AI applications trained to execute specific tasks (Brynjolfsson & McAfee, 2017). Still, the usage of AI has had a significant role in the rise of some of the most successful companies like Apple, Alphabet (parent company of Google), and Amazon (Bryson, 2019); hence the high expectations are entitled to some extent.

It is predicted that the effects of AI will be magnified in the coming decades when AI applications are implemented in various industries (Brynjolfsson & McAfee, 2017). That will force the companies to transform their core processes and business models. To stay in the competition, companies today are deploying AI systems to be more efficient. Based on Brynjolfsson and McAfee (2017), "Over the next decade, AI won't replace managers, but managers who use AI will replace those who don't".

2.1.1 Machine Learning

As the most common form of Artificial Intelligence today, machine learning has been coded to learn either by human supervision or by its own with training data. By the definition of Alpaydin (2014, p. 1-2), machine learning refers to a computer program that is programmed to optimize its performance by using example data or past experience. To learn and understand the provided data set, the machine learning model applies different algorithms. Machine learning models can be used to make future predictions or to gain knowledge from the past. If machine learning methods are applied to large databases, it is called data mining. (Alpaydin, 2014, p. 1-4).

The types of machine learning are determined by how feedback is used in the training process. The three main types are unsupervised learning, reinforcement learning, and supervised learning. In supervised learning, the machine has training data with test examples consisting of inputs and outputs, and the machine learns a function that maps inputs to outputs. Reinforcement learning the model is taught with rewards and punishments. The correct outputs are not provided, but the feedback is given after the machine provides the output. In unsupervised learning, the expected inputs or outputs are not provided, and the feedback is not explicit. The machine learns by detecting patterns in the training data. One of the most common tasks for unsupervised learning is clustering, which means recognizing patterns from the unlabeled data set. (Russel & Norvig, 2010, s.694-695).

Historically in computer science, the emphasis has been on developing better algorithms. However, within the last decades, the interest has shifted more to collect and create usable data (Russel & Norvig, 2010, p. 694-695). To train a machine learning model, the data is the key. Even though the amount of data is growing at exponential speed, the major challenge is the usability of the data, as the raw data is unlabeled or unstructured and requires much effort for refining. To create more powerful machine learning models, the solution is not a new specific algorithm, but the usable example data and sufficient computing power. (Alpaydin, 2016, p. 16-17).

Techniques like deep learning can be used as part of a solution, as deep learning requires a smaller training data set. A deep learning model can be fed with raw data, and it can be used for detection and classification. The models using unsupervised deep learning are expected to become more critical in the future. Deep learning can be used for more complex tasks like natural language understanding and imitation of human vision, and in the future combine it with complex reasoning. (LeCun, Bengio & Hinton, 2015).

Besides AI and machine learning, there are a couple of other essential concepts to understand this paper's research area. When talking about AI, people often think about robotics, bots, and autonomous machines. Robotics refers to AI's embodiment, and bots refer to virtual entities, usually powered by machine learning, such as virtual assistants or chatbots. Autonomous machines, such as vehicles, robots, or production machinery, differ from automation with the capability to learn and make decisions fully or semi-autonomously without human supervision.

2.2 AI Ethics

Due to the AI systems' capability to learn and make decisions autonomously and the broad interest to deploy AI in various fields, the interest and need for ethical research and guidelines have increased. In academia, the discussion and research of AI ethics have been running for decades, but it rarely crosses with the development of AI systems (Vakkuri & Abrahamsson, 2018). The research of AI ethics has been focusing on the potential of AI on a theoretical level and on finding technological solutions, even though often a broader perspective is required (Brundage, 2014). AI ethics is a continually evolving research area that is interesting for several domains like computer science, economics, 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 (Vakkuri & Abrahamsson, 2018). AI ethics is also important from a societal perspective, and institutions like the European Union are putting effort to establish ethical guidelines of AI usage. Also, for private organizations, AI ethics is a concerning issue, as they are responsible for the acts of the incorporated AI systems.

Ethics (also called moral philosophy) is a research area of philosophy that aims to define the concept of right and wrong and resolve questions of human morality. Ethics is divided into three subject areas:

1. Metaethics that investigates the origin of ethical principles.
2. Normative ethics with a more practical viewpoint to determine a moral course of action.
3. Applied ethics examines controversial issues in domain-specific situations aiming to determine the obligated or permitted actions. (Fisher, 2020).

Applied ethics include environmental concerns and human rights (Fisher, 2020), and it often concerns real-life situations that require quick decision making (Ala-Pietilä et al., 2019). One sub-field of applied ethics is computer ethics that includes AI ethics, which involves the ethical issues raised by the development, deployment, and use of AI (Ala-Pietilä et al., 2019).

Computer ethics studies the moral questions associated with the development, application, and use of technology (van den Hoven, 2009). Computer ethics has its roots in the 1940s, but the subject boomed in the late 1970s when the first significant problems, like computer crimes and invasions of privacy, became public concerns (Bynum, 2001). During 1990, computer ethics merged with information ethics that studies the moral questions connected to the availability, accessibility, and accuracy of informational resources (Floridi, 2009). Within the last three decades, the field of computer ethics has grown radically, and it is assumed to gain further importance in the future, as technology is becoming more and more globally significant and ultimately an undivided part of people's everyday lives (Bynum, 2001).

Besides computer ethics, which focuses on how humans use computers, the other important sub-field of AI ethics is machine ethics. According to Moor (2006), machine ethics is interested in moral embedded into machines. The core concept in machine ethics is ethical agents that can make ethical decisions. An average adult human is a full ethical agent. A machine can be seen as an ethical impact agent that has an ethical impact to its surroundings, or as an implicit ethical agent that is coded to follow a particular ethical framework in executing a specific task, or as an explicit ethical agent that makes ethical decisions in complex fast-changing situations (Moor, 2006). A machine that could behave like a full ethical agent would probably require the development of Artificial General Intelligence, AGI, that refers to the level of intelligence comparable with human intelligence, or superintelligence that refers to machine surpassing human intelligence.

2.2.1 Principles of AI ethics

The ethics of AI is often defined by using a list of principles, laws, or guidelines for AI developers or implementors to follow. Often, in the base of ethics of AI is the reference to Isaac Asimov's (1942) imaginary laws in science fiction literature:

1. The robot must not harm or endanger humans
2. The robot must obey the human command unless the command conflicts with the first law.
3. The robot must protect its existence unless it conflicts with laws 1 or 2.

In this research's scope, it is not interesting to focus on the ethical problems of the future, such as the construction of the moral code of the conscious machine, but on the challenges that are encountered today. It is yet relevant to understand the base and roots of AI ethics.

Jobin et al. (2019) mapped the corpus, including the grey literature, such as corporations' white papers and reports, of AI ethical guidelines and principles revealing the 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 was interpreted as explicability, understandability, interpretability, communication, disclosure, and showing. Justice was most often interpreted as fairness, consistency, inclusion, equality, equity, (non-)bias, and (non-)discrimination. Most often, non-maleficence referred to general security, safety, and causing of foreseeable or unintentional harm. Responsibility and accountability referred to liability and integrity, or to the different actors named as accountable for AI's actions. Privacy in AI ethics means both a value to uphold and a right to be protected. (Jobin et al., 2019).

The most frequent requirement in the AI ethics literature was transparency, followed by justice and fairness (Jobin et al., 2019). Transparency and fairness are required to ensure the system's ethical function. Without transparency, fairness cannot be evidenced in the system. A third, closely connected issue is accountability. Together these three constructs the FAT (fairness, accountability, and transparency) theorem.

2.2.2 AI Ethics in Practice

Within the last years, the questions about responsibility and transparency in autonomous systems have been visible in mainstream media due to pedestrian fatalities with self-driving cars. In situations like that, it might be challenging to detect why the mistake occurred and who is responsible: the driver, the car developer, or maybe the pedestrian themself? Humans design AI systems and, therefore, it is a matter of human responsibility (Bryson, 2019); hence the car itself cannot be responsible for an overrun. In these situations, transparency of the system is required to fix the system and prevent future accidents.

Autonomous driving is a broadly discussed topic in the AI ethics field. It has opened the venue to non-practitioners to join the conversation and understand the issues related to AI ethics. MIT's research Moral Machine collected 40 million answers to their online experiment, which studied the decisions in ethical situations related to autonomous driving (Awad, Dsouza, Kim, et al. 2018). The discussion around autonomous vehicles and autonomous driving has saturated, and it might take the focus away from more relevant issues. Still, during the last years the discussion around AI ethics has opened to concern a broader scope of topics.

Cathy O'Neil's popular book, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016), brought algorithmic inequality and biased algorithms to a broader audience outside academia and data science fields. The book showcases problems, especially in US legal and public systems. Racial factors can determine the futures of mortgage applicants or convicted criminals, even if racial information is not accessible to the algorithm (O'Neil, 2016). One commonly known discriminative case was Amazon's AI recruiter, who preferred male applicants in technical positions due to the historical data and dominance of men in technical roles (Dastin, 2018). If the systems are not transparent, discrimination and biased decisions cannot be tracked and fixed.

Regulators like the European Commission are increasingly interested in the topic. In 2018, European commission assembled a High-Level Expert Group on Artificial Intelligence, AI HLEG, with the core purpose to support the implementation of the European Strategy on Artificial Intelligence. The commission's vision is to increase investments in AI, prepare for socio-economic change, and ensure an appropriate ethical and legal framework. European Commission's AI HLEG (2019) has identified 'Trustworthy AI' as the EU's foundational ambition for ethical AI. Trustworthy AI has three components, each of them necessary but not sufficient in achieving Trustworthy AI. The AI system should be *Lawful*: compliant with all applicable laws and regulations, *Ethical*: ensure adherence to ethical principles and values, and *Robust*, from a technical and social perspective, because even with good intentions, AI systems can cause unintentional harm. Even though it is desirable to have all three components working in harmony, it is not always possible in real life. (Ala-Pietilä et al., 2019).

World Economic Forum (WEF 2016) has clustered the open questions in AI ethics with the following categories:

1. Rising of unemployment due to job loss for machines
2. Inequality of incomes and whether the use of AI increases the concentration of incomes
3. Humanity and AI's effect on human behavior and interaction
4. Protection of errors and flaws in AI systems
5. If the use of AI magnifies the human biases
6. Protection of AI systems from malicious actors
7. Avoidance of unwanted side effects
8. Potential singularity and protection against powerful machines
9. The machine's rights for conscious beings.

This paper focuses on today's issues connected to the explainability and understandability of AI algorithms and algorithmic decision making. These issues are connected to points 4, 5, and 7.

Companies and private organizations are also establishing their ethical frameworks and principles. In 2019, the Finnish governmental initiative The Age of AI released a challenge for AI's ethical development. Seventy companies participated in the challenge including many of Finland's largest corporations (Ministry of economic affairs and employment of Finland, 2020). Large practitioner organizations, such as Google, Intel, and Microsoft, have also presented their guidelines concerning ethics in AI (Vakkuri, Kemell, and Abrahasson, 2019).

In academia, guidelines and principles aim to structure the research field. One notable example is IEEE guidelines for Ethically Aligned Design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). In 2018 two closely topic-related conferences were launched; *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, that gathers researchers and authors from different disciplines to elaborate on the impact of AI on modern society (AAAI/ACM, 2017) and *FAT* conference* that gathers a diverse community of scholars to tackle the issues with algorithmic fairness, accountability and transparency in socio-technical systems (ACM FAccT Conference, 2020). Here FAT refers to the fairness, accountability, and transparency theorem that was mentioned earlier.

The frameworks' challenge is that they tend to lack the practices to implementing them into practice and require more work to be production-ready (Morley et al. 2019). The principles and guidelines are a good starting point for ethical discussion but, unfortunately, the principles presented in the literature are not actively used in practice (Vakkuri, Kemell, Kultanen, and Abrahamsson, 2020). This paper investigates the current research corpus with empirical evidence to understand the AI ethics research field closer to real-world issues. The interest is in transparent systems, one of the key challenges in AI ethics in practice (Jobin et al., 2019) and governance (Ala-Pietilä et al., 2019). Transparency is investigated together with fairness, as fairness often requires transparency from the system. The next chapter provides the background of transparency and explainable AI systems.

2.3 Explainable AI

Machine and deep learning techniques are used to automate decisions for better or faster decision-making processes. Unfortunately, the use of complex techniques, such as deep learning, makes the decisions hard to understand for humans. To ensure the right for explanations, legislation, such as GDPR, is permitting individuals a right for a meaningful explanation for decisions made by automated systems. Explainable AI (XAI) refers to an AI system that can explain its decisions. (Schneider & Handali, 2019).

The AI models are expected to be interpretable, which means that it can explain the decision in understandable terms to a human (Holm, 2019). A sophisticated knowledge extraction and preference elicitation is required to extract a meaningful explanation from the raw data used in the decision process (Schneider & Handali, 2019). This often means that a tradeoff must be made between accuracy, effectivity, and interpretability (Adadi & Berrada, 2018).

Interpretability is merely not just a technical problem. To gain interpretability of machine learning systems, it is required to focus on humans, rather than technical aspects, and provide personalized explanations for individuals (Schneider & Handali, 2019). Understanding of human decision-making and explanation-defining provides a good ground for XAI. That requires multidisciplinary collaboration and the use of existing research from social sciences such as philosophy, psychology, and cognitive science (Miller, 2018).

Besides social science and artificial intelligence, the scope of XAI includes Human-Computer Interaction, which studies the relationship between humans and machines. More precisely XAI is only one of the challenges in the scope of Human-Agent Interaction, which studies the relationship between humans and AI powered machines. The problem sphere of collaboration and interaction between humans and exponentially developing thinking machine agents is much greater than just the challenges with interpretability. (Miller, 2018).

Interpretability might not be expected from AI systems that do not have significant consequences of a wrong decision or if users trust the system even if it is known to be imperfect (Holm, 2019). For example, if a non-interpretable AI, like a world-famous AlphaGo, can beat a human in the Go game, the explanations of the tactical game decisions are not important (Samek, Wiegand, & Müller, 2017). Or if the AI system detects cancer cells, perhaps the system's benefits are larger than the potential pitfalls caused by a lack of interpretability.

In many cases though, interpretability is required. The reasons for the need for XAI vary. Based on Wachter, Mittelstadt, and Russell (2018) the reasons might be:

1. to inform the subject of the reasoning of a particular decision, explain the reasons for rejection, or
2. to understand how the decision-model needs to be changed to receive the desired decisions in the future.

Of course, the application area and purpose impact the need for interpretability.

Explainable and understandable systems are required for society to trust and accept the algorithmic decision-making systems (Wachter, Mittelstadt, and Russell, 2018). Better explanations can also improve existing models and open new opportunities, such as the use of machines for teaching humans (Schneider & Handali, 2019). 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 (Samek et al.,2017). Explainability is also important when assigning responsibility in case of a system failure (Samek et al., 2017), such as in the case of an overrun of a self-driving car.

Explainability is essential and beneficial yet challenging a challenging task. To understand the overall topic of XAI, other concepts are needed to explain. The following chapters tell about transparency, the black box problem, and algorithmic bias, which are closely connected to XAI. The last chapter of the background study concludes the literature review and justifies the motivation for Systemic Mapping Study.

2.3.1 Transparency

Both the EU AI Ethics guidelines (AI HLEG 2019) and EAD guidelines (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019) consider transparency an essential ethical principle. Even though transparency is named as one of the primary principles of AI ethics (Jobin et al., 2019), actually transparency can be seen as the pro-ethical circumstance, which makes the implementation of AI ethics possible in the first place (Turilli and Floridi, 2009). Without understanding how the system works, it is impossible to understand why it malfunctioned and, consequently, establish who is accountable for the malfunctions' effects.

The meaning of transparency varies depending on the subject, which makes the concept vague and misinterpretations likely. In the discipline of information management, transparency often refers to the *form* of information visibility, such the access to information. In computer science and IT disciplines, transparency often refers to a *condition* of information visibility, such as computer application's transparency to its user, and how much and what information is made accessible to a particular user by the information provider. In this paper, the term transparency is used in the meaning of the condition of information visibility. (Turilli and Floridi 2009).

Even though transparency is often required, the issue is not that simple. The information provider (e.g., companies or public institutions) must define who has the right to access the information and accessibility conditions (Turilli and Floridi 2009). Legislation, such as GDPR, might control the access and sharing of a specific type of information between users. Especially in medicine and health, the full transparent access to patient's data across the organization or the country borders could accelerate the speed of development, such as drug discoveries. However, in the other hand it could lead to ethical challenges and misuse of highly sensitive personal data.

Instead of seeing transparency as an ethical principle, it would be more accurate to treat it as an ethically enabling or impairing factor, the 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 with e.g., discrimination, privacy, and security. (Turilli and Floridi 2009).

2.3.2 Black box problem

It is called a "black box" when the AI model is not understandable and cannot provide a suitable explanation for its decision (Adadi & Berrada, 2018). A black box refers to a model that is either too complicated for any human to comprehend or proprietary to someone (Rudin, 2019). Typically, deep learning models belong to the first category. To understand the black box, the model needs to be built to be *interpretable* or create a second model that explains the first black-box model (Rudin, 2019). Interpretability in AI context refers to the capability to understand the overall work logic in machine learning algorithms, not just the answer (Adadi & Berrada, 2018). The terms interpretability and explainability are often used as synonyms (Adadi & Berrada, 2018), which can be challenging as the level of required understandability is different. In the public discussion, the term Explainable AI is more used than Interpretable AI, whereas in academic discussion, the situation is the opposite (Adadi & Berrada, 2018). Current AI regulation, such as GDPR, requires the right to explanation, not an interpretable model, which might cause problems in certain areas (Rudin, 2019).

A second post-hoc explainable model might provide explanations that do not make sense or that are not detailed enough to understand what the black box is doing. If the provided explanation would give a full understanding of the model, that would make the system interpretable. Secondary explanatory models are often not compatible with information outside the black box. The lack of transparency in the whole decision process might prevent the interpretation by human decision-makers. Secondary models can also lead to overly compilated decision pathways when the transparency is required actually from two models: the original black box and the explanatory model. (Rudin, 2019).

Neither of the interpretable machine learning models is challenge-free. First, because it is a computational challenge to build one. Second, the AI system's total transparency can jeopardize the system owner's business logic, as the system owner must give out part of their intellectual property. Constructing the interpretable model is often expensive as it requires domain-specific knowledge, and there are no general solutions that would work in different use cases. Creating an interpretable model is a challenge to find the balance between interpretability and accuracy, as interpretable models tend to reveal hidden patterns in data, which are non-relevant to the subject. (Rudin, 2019).

2.3.3 Accountability and Algorithmic Bias

Besides 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 (Adadi & Berrada, 2018). Responsible AI consists of three main pillars: *transparency* (described in the previous chapter), *responsibility* which requires "to link the AI system's decisions to the fair use of data and to the actions of stakeholders involved in the system's decision", and *accountability*, which requires that the "decisions must be derivable from, and explained by, the decision-making algorithms used" (Dignum, 2017).

Accountability refers to an actor who is accountable for the decisions made by AI. To establish accountability, the system must be understandable. The lack of transparency and accountability of predictive models can cause serious problems, such as discrimination in the juridical system, endangering someone's health, or misuse of valuable resources (Vakkuri, Kemell, Kultanen, and Abrahamsson, 2020). One of the recent incidents with the lack of transparency and accountability was an algorithm used to determine the final grades for International Baccalaureate students. The grades were inconsistent and worse-than-expected, which harmed the university selection of the individuals (Evgeniou et al., 2020).

Based on Vakkuri et al. (2020) 's research, transparency is the enabler for accountability, and together transparency and accountability motivate the responsibility. Finally, responsibility produces fairness. The fairness is often linked with algorithmic biases. AI system might repeat and magnify biases in our society, like to segregate groups with a history of discrimination, such as 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" (Mitchell, 1980). Machine learning systems are neutral and do not have opinions, but the models are not used in voids, which makes them vulnerable to the biases of humans. The reason for discrimination and unfairness with machine learning 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 might reveal biases invisible during the development process. There is no easy solution to ensure fairness of algorithmic decisions. (Vaele and Binns, 2017).

Vaele and Binns (2017) identified three distinctive approaches to ensure fairer machine learning. First is the third-party approach, where another organization is managing data fairness. Second is the collaborative knowledge base approach, where linked databases containing fairness issues are flagged by researchers and practitioners. A third approach is an exploratory approach, where exploratory fairness analysis is performed to the data before training the model or before the practical implementation of the model.

In this paper, the interest is in the exploratory approach because it is connected to the black box problem (Vaele and Binns, 2017). In this paper, the biases are studied from XAI's perspective, which aims to bring transparency to the

AI system. Less emphasis is dedicated to research on how data collected or processed to avoid biases.

2.4 Conclusion of Background Study

The research of AI ethics lacks harmony and standard agreement on defining the core principles (Jobin et al., 2019). This paper aims not to solve the issue of definitions for fairness and transparency but instead to investigate the existing research connected to transparency as understood in this paper, a requirement from the AI system to provide an understandable explanation if required in the context of the application. This requirement applies to systems that are non-explainable due to the training method or biased due to training data. This paper takes no stand upon ranking the principles. Instead, it aims to provide a more in-depth understanding of one of them: transparency.

The research field of XAI studied as a subfield of AI ethics, is researching the challenges and looking for a potential solution for transparent machine learning models, and therefore enable the fulfillment of ethical principles such as accountability, responsibility, and fairness. XAI can benefit a broad range of domains relying on AI systems. Especially in domains such as legal, finance, military, and transportation, the need for XAI is emphasized (Adadi & Berrada, 2018). In such domains, the AI systems are in direct influence on people and can cause injuries (Adadi & Berrada, 2018). In other domains, transparency might not be required. There is no one-for-all framework or solution available for transparency issues; hence the domain-specific solutions and frameworks are required.

The research field is short of the knowledge of industrial practice's current state with AI ethics (Vakkuri et al., 2020). Rudin (2019) is concerned that the XAI field suffers from the distancing of real-world problems. Based on Rudin (2019), the recent work in the field is more concerning the explainability of black boxes than the interpretability of the model. On the other hand, Adadi and Berrada (2018) were concerned that interpretable machine learning takes all the attention and leaves other promising explainable models under-explored. Their research also showed that XAI's impact is spanning in a broad range of application domains. However, the lack of formalism regarding problem formulation and clear, unambiguous definitions burdens the research field. Besides, they noted that the human's role is not sufficiently studied. A recently published paper recognized the same challenge with the lack of user-centric design in XAI (Ferreira and Monteiro 2020). Došilović et al. (2018) stated that XAI is a complex study field lacking common vocabulary and formalization.

AI ethics and XAI are broad, versatile topics with increasing importance. This paper aims to give a holistic view of the research field through a profound literature review. It is required to understand what is studied in AI ethics research to understand the role of explainable AI. More systemic research is required for that purpose, and in the next chapters, Systemic Mapping Study is used to understand the study field of AI ethics and how XAI is manifested in the research.

3 LITERATURE SEARCH FOR PRIMARY STUDIES

The literature review is conducted by using a systematic mapping study (SMS). The SMS continues an SMS of Vakkuri and Abrahamsson (2018) that studied the AI ethics research field's key concepts. In this paper, the existing dataset was complemented with the latest research. The existing dataset included the papers from 1/2012-7/2018. Vakkuri's and Abrahamson's (2018) goal was "to identify and categorize keywords used in academic papers in the current AI ethics discourse and by that take first steps to identify, define and compare main concepts and terms used in discourse." Their goal is aligned with this paper's goal of identifying the role of explainable AI in the research field of AI ethics. After the primary search, the datasets were combined to a database for which the following process steps were performed.

The research area of the Ethics of Artificial Intelligence is emerging. Due to the research area's emerging nature, this literature review is done cumulatively to better understand the state of research. The primary goal for cumulative review in Information Systems is to evaluate and understand the size and scope of existing literature (Templier & Paré, 2015). As the research area is fragmented across various domains and databases, the cumulative research approach offers tools, such as thematic analyzes, to understand the data and summarize the prior research material (Templier & Paré, 2015). In this paper, the cumulative research approach is made by conducting a Systemic Mapping Study, SMS.

The main focus for SMS is to "provide an overview of a research area, and identify the quantity and type of research and results available within it" (Petersen, Feldt, Mujtaba & Mattsson, 2008). SMS is traditionally used in medical research, but it has become a popular study method in Information Technology (Budgen, Turner, Brereton & Kitchenham, 2008). SMS is well suited in situations where the research area and topics are more open than traditional systemic literature reviews. These fields might lack high-quality primary studies (Budgen et al., 2008). SMS gives an overview of the research topic, and later it can be complemented with a systematic literature review to investigate the state of evidence in a specific focus area (Petersen et al., 2008). There are high-quality studies about AI ethics, but the research field is fragmented under different domains. AI technologies are emerging, which leads to constant change with ethical concerns. The SMS can give structure and help to conceptualize the research area.

SMS suits well situations in which a particular research area is studied from a new perspective. For this paper, the SMS results are analyzed to understand the role of explainable AI in AI Ethics literature, and what topics are connected to explainability. In the following chapters, SMS methodology and the process of the literature search are explained and visualized.

3.1 Defining the research question and the research process

The SMS aims to identify the potential research gaps and trends, including the understudied topics and research types. The expected outcome is "an inventory of papers on the topic area, mapped to a classification" (Petersen et al. 2015). The research question defines the scope of the research and sets the goals for the research. Typically, the main goal of an SMS is to create an overview of a particular research area and identify and visualize the quantity and type of research and results available. The research questions should reflect those goals. (Petersen et al., 2008).

From the perspective of this paper, SMS's goal is to understand which ethical concerns are covered in AI literature and to analyze the topics connected to explainable AI. This paper aims to understand the practical implementation and connection to real-world issues; hence the focus is on empirical studies. Based on Petersen et al. (2008), papers with the goal of 'Identify Best and Typical Practices' typically focus on analyzing empirical studies to determine the work in practice.

The research question for SMS can be quite a high level and cover issues such as what the addressed topics are, what empirical methods are used, and what sub-topics are sufficiently empirically studied (Kitchenham et al., 2011). This guideline forms the basis of the research question, "*What is the role of explainable AI in AI ethics' research field?*" divided into three sub-questions. The questions are:

[R1] What is researched in AI ethics research field with empiric evidence?

[R2] What is the current state of XAI in the research field of AI ethics?

[R3] What are the research gaps in the field?

The focus is to understand the coverage of XAI related topics and what are potential research gaps. It is first required to understand AI ethics' research area to answer the second research question [R2]. Hence the literature research is performed in the AI ethics research field. More profound analyses, classification, and mapping are performed only to papers related to XAI.

The processes of building SMS is cumulative, and it includes several rounds of screening the papers. The process steps and outcomes are presented in Figure 1. The headline of each block tells the process step, and the body reflects this research. The figure walks the reader through the whole study. The process model is based on The Systematic Mapping Process by Petersen et al., 2018.

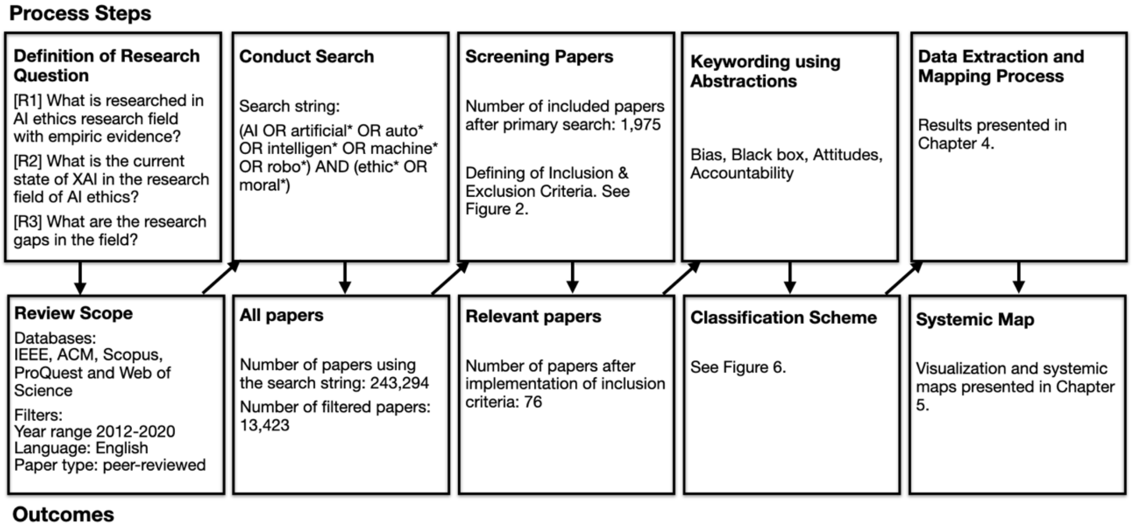

Process Steps
Definition of Research Question
[R1] What is researched in AI ethics research field with empiric evidence?
[R2] What is the current state of XAI in the research field of AI ethics?
[R3] What are the research gaps in the field?
Conduct Search
Search string:
(AI OR artificial* OR auto* OR intelligen* OR machine* OR robo*) AND (ethic* OR moral*)
Screening Papers
Number of included papers after primary search: 1,975
Defining of Inclusion & Exclusion Criteria. See Figure 2.
Keywording using Abstractions
Bias, Black box, Attitudes, Accountability
Data Extraction and Mapping Process
Results presented in Chapter 4.
Review Scope
Databases:
IEEE, ACM, Scopus, ProQuest and Web of Science
Filters:
Year range 2012-2020
Language: English
Paper type: peer-reviewed
All papers
Number of papers using the search string: 243,294
Number of filtered papers: 13,423
Relevant papers
Number of papers after implementation of inclusion criteria: 76
Classification Scheme
See Figure 6.
Systemic Map
Visualization and systemic maps presented in Chapter 5.
Outcomes

FIGURE 1 SMS Process based on Petersen et al (2018).

Because SMS's goal is not to give evidence, the quality of the chosen articles is not highly important, and articles are not evaluated based on their quality (Petersen et al., 2008). The articles do not need in-depth examination, so the number of articles included can be larger (Petersen et al., 2008). In this paper, the total number of papers included from five databases was 1975, and after applying the inclusion and exclusion criteria, the sample was narrowed to 76 papers. In the following chapters, each process step is further explained based on the theoretical framework.

3.2 Primary search

The first step in SMS is to identify the primary studies that contain relevant research results (Budgen et al., 2008). The search string and primary inclusion criteria were established in order to execute the literature search. As the literature search aims to find all the relevant papers, the inclusion criteria are not too narrow. The literature search identifies the primary studies using search strings on different scientific databases (Petersen et al., 2008). The literature search included a manual screening of databases to exclude papers that were not in this research scope but were shown in the search string results.

PICO (Population, Intervention, Comparison, and Outcomes) can be used as a guideline to develop a search string. The *population* refers to the main topic area researched, which in this paper refers to studies related to AI. The *intervention* refers to a topic that has an impact in the research area, which in this paper are topics connected to ethics and morals. There is no exact *comparison* in this study, as AI ethics is studied as a phenomenon. The search outcome is to understand the state of academic research related to the topic;

hence, only peer-reviewed papers were included. (Kitchenham and Charters, 2007).

This paper follows up the study of Vakkuri & Abrahamsson (2018), and the search strings and selected databases are adopted from their research. With the original research question of "*What topics are covered in AI ethics research?*" the search string consists of two parts: AI, and its synonyms (robotics, artificial, intelligence, machine, and autonomous) and Ethics and its synonyms (moral). The final search string is:

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

Alternatively, split into three search strings that were required for IEEE:

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

The search was narrowed to conclude only the headline and the abstract to find papers that focused on AI ethics. The databases, search strings, and search results from 2018-2020 are presented in Table 1. The Table shows the total papers found with the search string, the papers after applying the filters, related papers that met the criteria of inclusions, and finally, the included papers that present the number of papers included per database after deleting the duplicates per database.

TABLE 1 Search Results 2018-2020

Results of primary search					
Database	**Search String**	**Total papers:**	**Filtered papers:**	**Related papers:**	**Included papers:**
IEEE Xplore	(AI OR artificial* OR auto* OR intelligen* OR machine*) AND (ethic*) (AI OR artificial* OR auto* OR intelligen* OR machine*) AND (moral*) (robo*) AND (ethic* OR moral*)	4247	938	413	280
ACM Digital Library	(AI OR artificial* OR auto* OR intelligen* OR machine* OR robo*) AND (ethic* OR moral*)	1227	579	457	457
Scopus	(TITLE-ABS-KEY (ai OR artificial* OR auto* OR intelligen* OR machine* OR robo*) AND TITLE-ABS-KEY (ethic* OR moral*))	51142	6,029	1457	1449
ProQuest	noft((AI OR artificial* OR auto* OR intelligen* OR machine* OR robo*)) AND noft((ethic* OR moral*))	172296	2,144	198	198
Web of Science	(TS=((AI OR artificial* OR auto* OR intelligen* OR machine* OR robo*) AND (ethic* OR moral*)))	19,856	3,775	563	543
Totals		**248768**	**13465**	**3088**	**2927**

As SMS screens a large number of papers, the selection of databases is essential. The search is performed in five electronic databases. Databases represent the two central databases of information system science: IEEE and ACM, and three large multidisciplinary databases Scopus, ProQuest, and Web of Science. In total, there were 248,768 results in the five databases. For the literature search, the inclusion criteria consist of three filters; publication year (2012-2020), document type (peer-reviewed articles and proceeding papers), and language (English).

Due progression in the development of AI in early 2010, the research that has been done before 2012 is often nonrelevant today. The field has changed radically due to the invention of deep learning and other modern AI techniques. Only the years between 2012-2020 are interesting for this research. As the research continues the work started in 2018, only the missing years 2018-2020 were now included in the study. After the literature search, the extraction of papers from 2012-2018, and the extraction of papers from 2018-2020 was compounded. This paper presents only the literature search results of the year range 2018-2020.

The search with three filters (document type, publication year, and language) performed in five databases IEEE, ACM, Scopus, ProQuest, and Web of Science resulted in 13,465 papers. All the resulted papers were screened manually during May and June 2020. The effectivity of the filters is presented in Table 2. The numbers indicated the number of papers in a column. The filters were always applied in the same order; document type, year, language; hence the language column shows the final number of papers after applying all three filters.

TABLE 2 Effectivity of Applied Filters

Effectivity of Applied Filters			Filter		
Date	**Database**	**Before Filters**	**Document type**	**Year**	**Language**
5.5.2020	IEEE (search string 1)	1724	1619	405	405
5.5.2020	IEEE (search string 2)	1624	1581	328	328
6.5.2020	IEEE (search string 3)	899	808	205	205
15.5.2020	ACM	1227	779	579	579
5.6.2020	Scopus	51,142	35,847	6,654	6,029
12.6.2020	ProQuest	172,296	11,352	2,377	2,144
19.6.2020	Web of Science	19,856	16,987	4,392	3,775

In manual screening, the papers that did not meet the inclusion criteria were excluded. For example, papers that examine the use of AI systems to fix a particular ethical problem, such as detecting fake news in social media, were excluded from the study. In this paper, the interest is in ethical questions related to the use of AI. The manual screening was performed only to the abstracts. The final numbers of papers after each process step are presented in Table 3.

TABLE 3 Total Papers Included in Different Process steps

Number of papers after each search process steps	Papers
Results with the search string	243,294
Filtered papers	13,423
Manually included papers	3088
After deletion of duplicates in separate datasets	2927
After deletion of duplicates cross datasets	1975

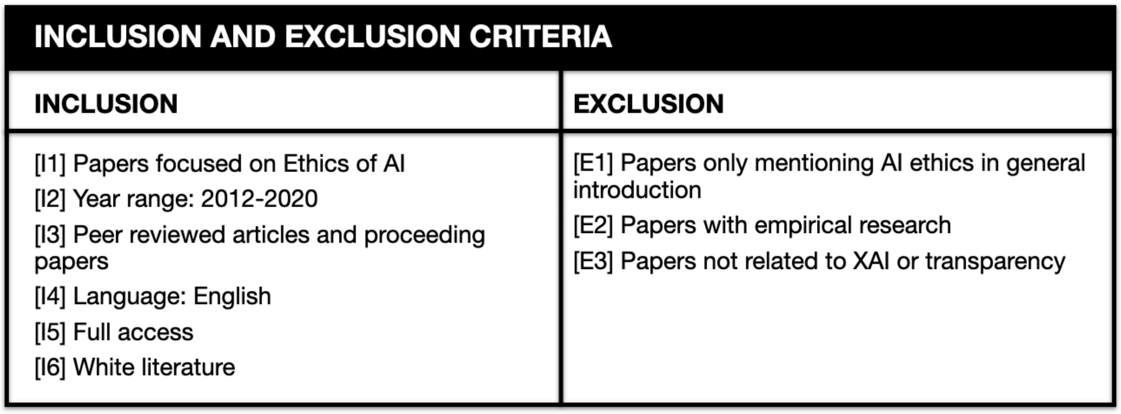
Manual screening narrowed the scope into 3088 papers. After screening each database, the duplicates were deleted, which narrowed the scope to 2927 papers. Next, the included papers from the separate databases were conducted into one database, and duplicates were excluded, resulting in a total of 1975 papers included in the second round of the SMS process.

3.3 Inclusion and Exclusion

The primary search resulted in 1975 papers. The second step of SMS is to examine the selected papers and find the primary studies (Budgen et al., 2008). This process requires defining more narrowing inclusion criteria. The reason for the inclusion criteria is to exclude studies that are not relevant for answering the research question (Petersen et al., 2008). A paper needs to fulfill all inclusion requirements and none of the exclusion requirements to fit the inclusion criteria.

The inclusion process is guided by the research's goal and the desirable contribution (Paternoster, Giardino, Unterkalmsteiner, Gorschek, Abrahamsson, 2014). This study aims to map the relevant research area of ethics of AI in the domain of Information system science; hence only the papers focusing on the ethics of AI (I1) were included. Due to a large number of included papers after the primary search, it was decided to include only the papers with full access (I5). Papers without access to full texts can be excluded from SMS (Petersen et al., 2008). 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 quality of the included papers, only white literature, papers were included (I6). White literature refers to full papers published in venues of a high control and credibility, and it excludes pre-prints, technical reports, blogs, and other types of publications that are referred as grey and black literature (Garousi, Felderer, and Mäntylä, 2019).

In SMS studies, exclusion criteria might include 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 (Petersen et al., 2008). The first exclusion criteria (E1) is the exclusion of papers that are not contributing to the AI ethics research and only mention the potential ethical issues related to AI in the general introduction. Moreover, in this paper, the interest is in practical AI implementation rather than a philosophical concern; hence, the papers without empirical research were excluded from the study (E2). In the final screening, papers not focusing on XAI or related topics were excluded (E3). Inclusion and exclusion criteria are presented in Figure 2.



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

FIGURE 2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria were established and defined during the screening process, and the JYU AI Ethic Lab supported the process. Inclusion criteria provided the general boundary and quality conditions, and exclusion criteria the more detailed limitations to distinguish the sample relevant for this paper. The purpose of the screening was to exclude papers that did not fit the inclusion and exclusion criteria. The inclusion and exclusion were done during several screening rounds to narrow the scope after each round and enable more in-depth inspection in the following round with a smaller sample. The process is further described in the next chapter.

3.3.1 Inclusion of academic papers with empiric research

For the first screening round of only three inclusion rules were applied; language [I4], access to full text [I5], and sufficiently used references [I6], which indicates the academic quality of the paper. The first round narrowed the total number of papers to 1532. After the first screening round, the dataset was combined with the dataset of 2012-2017 (n=403), which was separately screened to be equivalent to the dataset of 2018-2020. The combining of the datasets grew the database to 1935 papers. Later in this paper, it is only utilized the full dataset of 2012-2020 (n=1935).

The inclusion process is presented in Table 4 and later visualized in Figure 3. Table 4 shows the criteria applied in the screening round and the number of papers after the screening round. The screening rounds are closer described in this chapter.

TABLE 4 Screening Rounds in Inclusion and Exclusion Process

INCLUSION & EXCLUSION PROCESS

Round	Criteria	N
Screening 1: Review of reference list, language and paper type	Inclusion of Academic papers written in English and access to full text. Exclusion of workshop/keynote/newsletter introductions	1935
Screening 2: Review of abstract, keywords and scanning of full text	Inclusion of papers focusing on AI ethics, exclusion of papers mentioning or shortly describing issues related to AI ethics. Exclusion of non-peer reviewed papers like doctoral programs, course presentations and student posters	1065
Screening 3: Review of method chapter	Inclusion of papers with empirical data	212
Screening 4: Review of abstract, keywords and scanning of full text	Inclusion papers related to XAI, Exclusion: grey and black literature	76

In the next round, the papers were screened to exclude papers that did not focus on the Ethics of AI. Similar exclusion criteria have been used in other SMSes because the primary purpose of an SMS is to find the most relevant papers from a particular field (Vakkuri & Abrahamson (2018); Paternoster et al. (2014). During the second screening round, the quality of the paper was validated. Short papers, student works, course descriptions, workshops, or panel descriptions were excluded and papers that, for some other reason, did not meet the academic peer-review standards. The number of excluded papers is presented in Table 5, where is shown the number of excluded papers per inclusion and exclusion criteria.

TABLE 5 Excluded Papers

Excluded papers	
Reason of exclusion	**Number of papers**
No Full Access [5]	332
Not in English [I4]	9
Non-academic paper / grey literature [I3, I6]	311
Not focusing AI ethics [I1, E1]	734
No empirical evidence [E2]	852
Not related to XAI [E3]	133

The included papers were clustered into two categories; idea and data, to separate the empirical papers that are meaningful for this paper's goal. The empirical papers were manually separated during the screening, as it was the most liable way to ensure the sample of all the relevant papers. That would not have been possible if the primary search string contained the criteria for empiric material, as the search string would have become too complex due to all variations of search words needed to include. From the total 1065 papers that met the inclusion criteria, 212 papers were using empirical material. The "idea" papers consist of reports, philosophical papers, reviews, problem descriptions, and proposals. Systemic literature reviews were categorized into idea papers, as only primary research was included in this study.

In this paper, no further quality inspection was performed for the papers in the idea category. The idea category sample might include papers that are ideas, concepts, technical reports, or blog posts that are categorized as grey or black literature (Garousi, et al. 2019). Without any further examination, the ratio of idea papers represents 80% of the full sample. Future research is required to better understand the variety of theoretical papers better, but it is out of this paper's scope. After the exclusion of the idea category papers, the sample was 212 papers. The dataset (n=212) can be found in appendix 1.

3.3.2 Inclusion of high-quality papers focusing on Explainable AI

Next screening round the paper quality was examined, and grey and black literature papers were excluded from the sample. Exclusion of grey and black literature from SMS is crucial, as SMS aims to present the study field of AI ethics, and grey literature is often opinion or experience-based. Moreover, it might present results that are potentially prejudiced (Garousi et al., 2019). White literature was in the scope already in the preliminary literature search, which was limited to peer-reviewed papers. In previous screening rounds, other than academic papers were excluded. Excluded papers did not have sufficient references, were produced as student work, or were significantly short or narrow, such as workshop presentations and doctoral programs. In the third screening round, the papers were manually reviewed for grey and black literature. In the dataset of empirical data papers, this meant mainly excluding grey literature such as preprints, technical reports, lectures, data sets, and blogs (Garousi et al., 2019). At this stage, one working paper and two preprints were excluded from the sample.

The papers were skimmed and tagged based on the focus area to find the papers connected to XAI. As described in the background section, XAI is a vague concept, and there is no commonly agreed framework on what topics are considered to be included under the term. In the sample, XAI was the focus in 20 papers, but several papers focused on themes, such as transparency, that are closely connected to XAI. The papers focusing on responsible AI, algorithmic bias, or black box were included to ensure the inclusion of all relevant papers. In total, 76 papers were included in the final examination. The whole screening process is visualized in Figure 3, using similar visualities as in SMS of Belmonte, Morales, & Fernández-Caballero (2019). The inclusion side's sample size shows the included sample after each step and, in the exclusion side, the sample of excluded papers.

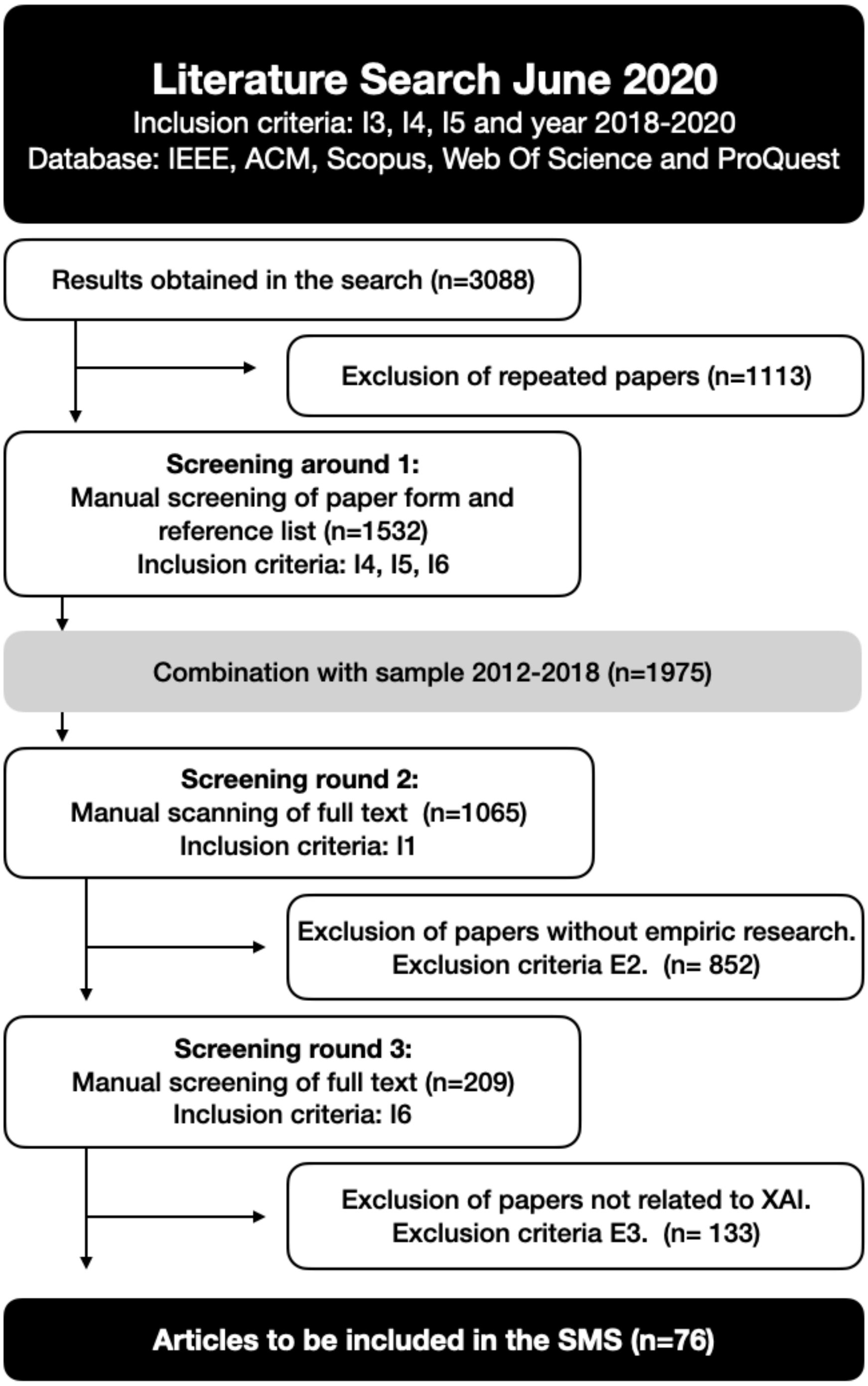

Literature Search June 2020
Inclusion criteria: I3, I4, I5 and year 2018-2020
Database: IEEE, ACM, Scopus, Web Of Science and ProQuest
Results obtained in the search (n=3088)
Exclusion of repeated papers (n=1113)
Screening around 1:
Manual screening of paper form and reference list (n=1532)
Inclusion criteria: I4, I5, I6
Combination with sample 2012-2018 (n=1975)
Screening round 2:
Manual scanning of full text (n=1065)
Inclusion criteria: I1
Exclusion of papers without empiric research.
Exclusion criteria E2. (n= 852)
Screening round 3:
Manual screening of full text (n=209)
Inclusion criteria: I6
Exclusion of papers not related to XAI.
Exclusion criteria E3. (n= 133)
Articles to be included in the SMS (n=76)

FIGURE 3 SMS Process Based on Belmonte et al (2019)

The final sample (n=76) included in SMS is further classified and analyzed in the following. A short analysis was performed to the sample of papers with empiric evidence (n=212) to understand the overall field of AI ethics. The analysis merely touched the surface of the full material, but a more profound investigation was out of the scope of this study.

3.4 Short analysis of AI ethics research field with empiric evidence

Future studies are required to understand the research area of AI ethics fully, but the short analyzes give sufficient background to mirror the role of XAI to the full sample of AI ethics research with empiric evidence (n=212). The empiric papers represent 20% of the whole sample of manually included papers (n=1065). This finding forms the first empirical conclusion.

> EC1: Most of the research papers in the field of AI ethics do not use empiric evidence. Only 20% of the papers provide empirical evidence.

Two dimensions were observed with the whole sample; emerging themes and the year of publication. The theme analysis was done during the keywording process described in the next chapter. A more profound analysis would require a more systematic approach.

As the research area is young and emerging, the year of a publication can provide insight into the research area's growth. The papers published per year are visualized in Figure 4. The size of the bar presents the number of papers published each year.

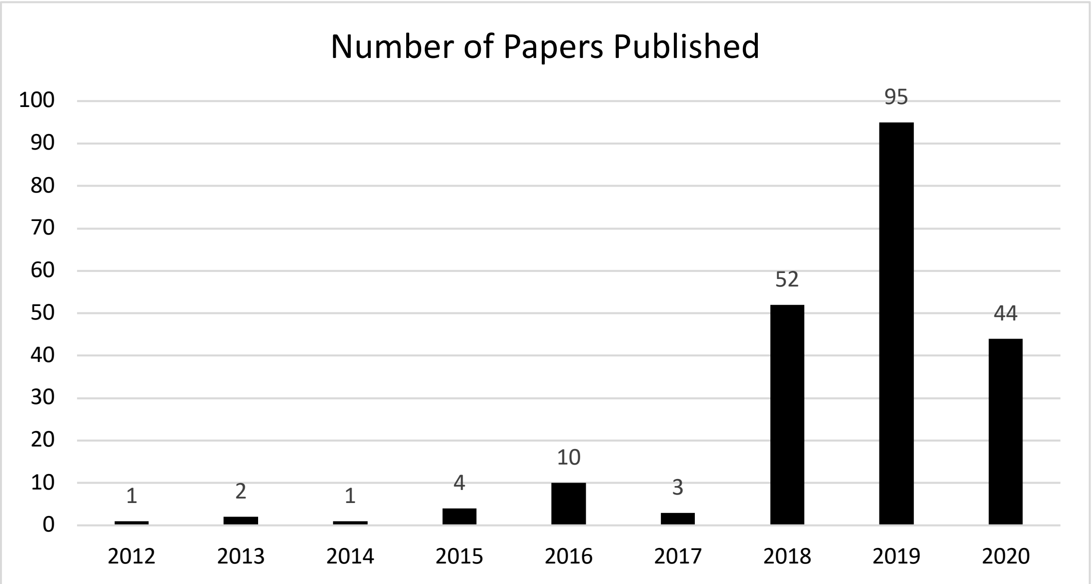


FIGURE 4 Annual changes in publication of empiric papers in AI ethics research area

The visualization reveals significant growth between the years 2017 and 2018. There is a clear correlation to public discussions as the discussion around AI ethics peaked in media in 2018 (Ouchchy, Coin, & Dubljević, 2020). This finding forms the second empirical conclusion.

> EC2: Empiric Research of AI ethics grew significantly in 2018,
> following the trend in public discussion.

Year 2018 also saw the launch of two large conferences, AIES and FAT*, that might have an impact on the topic. Overall, 54 papers, 25,8% from the whole sample (n=212) is published in AIES. The impact of FAT* is less significant, with only five papers published in there.

The papers related to XAI (n=76) represent 35,85% of the full sample of empirical papers (n=212). This finding forms the first primary empirical conclusion:

PEC1: Explainable AI is significant research focus on the study field of AI Ethics. From the empiric research papers published after 2012, 36.2% is related to Explainable AI.

A short analysis of themes was performed to the sample of 136 excluded empiric research papers. These papers were excluded due not having a relation XAI. The most frequent theme amongst these papers was Autonomous Vehicles that was the focus in 31 papers. Other frequent themes were Human-Robot Interaction (focus on 17 papers), Health/Care Robotics (focus on ten papers), Bots/Virtual Assistants (focus on seven papers), and Education (focus on six papers). Seventeen papers proposed a technical, mathematical, or design solution to solve an ethical issue in AI. Sixteen papers studied people's attitudes towards robots, autonomous vehicles, or other AI systems. As the inclusion of XAI did not require a full pertinence, the number of papers focusing on XAI are not comparable with other emerging themes. No further examination was performed to excluded papers.

4 CLASSIFICATION

During the final screening, papers were tagged based on the main focus of the paper. The main focus was deduced based on the keywords, abstract, and the headline of the study. Also, papers were tagged based on the research method and the type of data used in the study. This was an initial step towards a keywording process used to build a classification scheme, which helps arrange the papers into meaningful groups. Classification uses a systematic process where the classification schema is evolving and specifying during the process (Petersen et al., 2008). The classification process based on Petersen et al. (2008) is visualized in Figure 5.

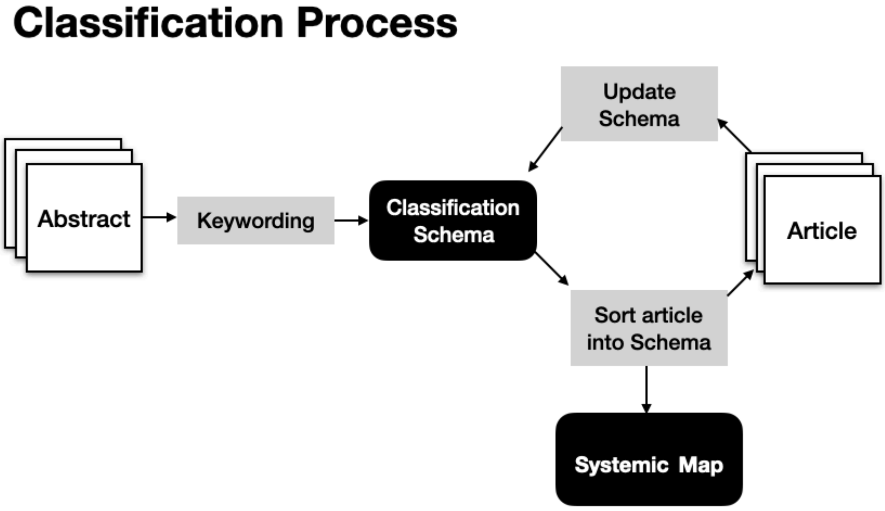


FIGURE 5 Classification Process Based on Petersen et al., 2008.

Keywording reduces the time required for building the classification schema and ensures that the classification schema represents the existing studies (Petersen et al., 2008). The first step of keywording was to identify the keywords, concepts, and the context of the research (Petersen et al., 2008). The process was started during the last stage of the inclusion process and continued with the final sample n=78 during the classification. The following chapters present the classification schema, classification results, and the overview of the final sample.

4.1 Classification schema

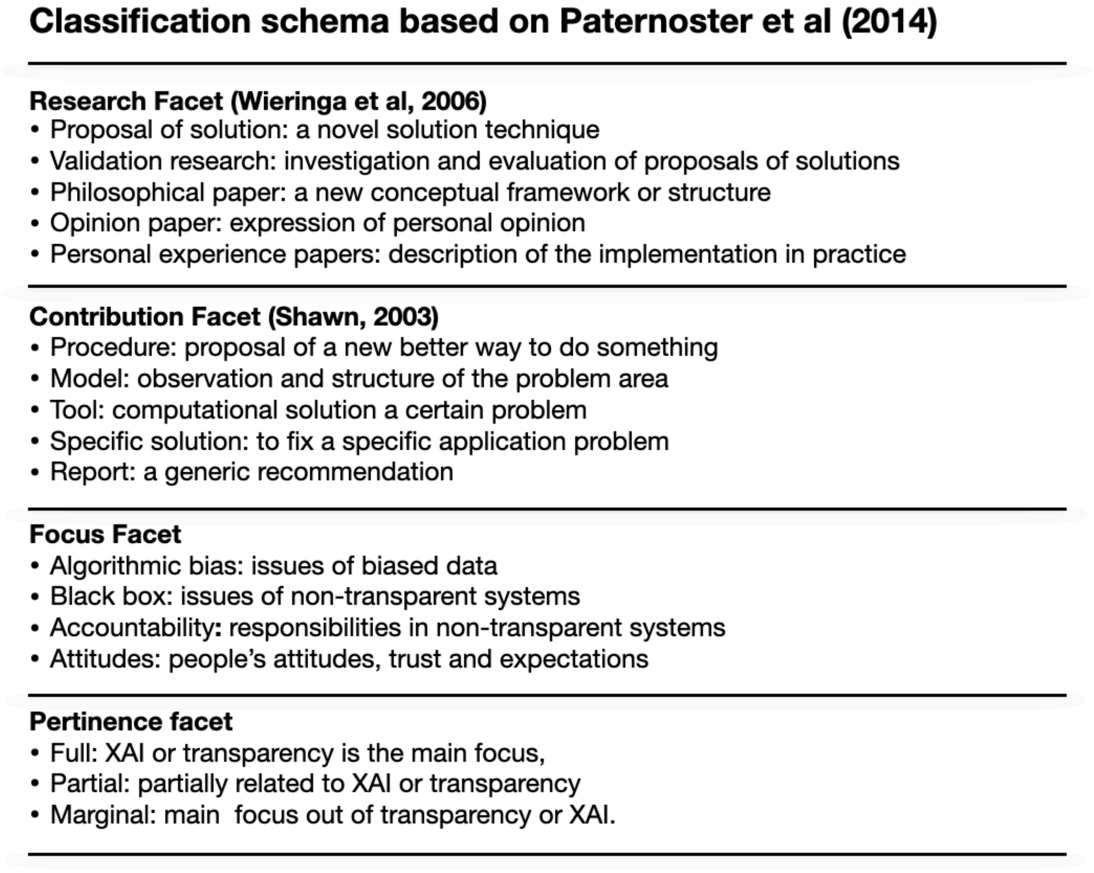
For the classification schema, the papers were examined in four facets adopted from SMS of Paternoster et al. (2014). The facets are research, contribution, focus, and pertinence. The facets are presented next. The summary of the classification schema is presented in Figure 6.

(1) Research facet. 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 argues for its relevance, without full justification, providing at maximum a narrow proof of concept. *Validation research* papers investigate the properties of their own 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. *Opinion papers*, expressing personal opinions without relying on related work or research methods. *Experience papers* are describing the implementation in practice, such as lists of lessons learned. The experience might be either the author's own experience or the experience of the person studied. (Wieringa, Maiden, Mead, and Rolland, 2006)

(2) Contribution facet. The aim is to identify the tangible contribution of the paper. That can be an operational *procedure* for development or analysis to provide a new, better way to do something, such as a design framework. Alternatively, a *model* representing the observed reality and structuring the problem area, or an implemented computational *tool* to solve a particular problem or a *specific solution* for a specific application problem. Alternatively, the contribution can be a piece of generic recommendation *advice* with a less systematic approach than the model. It often focuses on one example case and is more vaguely directive than the procedure. The contribution facet is based on the research of Shawn (2003).

(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,* the challenges with fairness due biased and discriminative training data or model, *black box,* challenges with non-transparent systems, and *accountability,* papers studying when and how the accountability of non-transparent system is divided. Some papers focused on understanding the attitudes, expectations, and trust towards non-transparent systems. These papers were categorized as *attitude.*

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



Classification schema based on Paternoster et al (2014)

Research Facet (Wieringa et al, 2006)

- Proposal of solution: a novel solution technique
- Validation research: investigation and evaluation of proposals of solutions
- Philosophical paper: a new conceptual framework or structure
- Opinion paper: expression of personal opinion
- Personal experience papers: description of the implementation in practice

Contribution Facet (Shawn, 2003)

- Procedure: proposal of a new better way to do something
- Model: observation and structure of the problem area
- Tool: computational solution a certain problem
- Specific solution: to fix a specific application problem
- Report: a generic recommendation

Focus Facet

- Algorithmic bias: issues of biased data
- Black box: issues of non-transparent systems
- Accountability**:** responsibilities in non-transparent systems
- Attitudes: people's attitudes, trust and expectations

Pertinence facet

- Full: XAI or transparency is the main focus,
- Partial: partially related to XAI or transparency
- Marginal: main focus out of transparency or XAI.

FIGURE 6 Classification Schema Based on Paternoster et al. (2014)

In all facets, the same paper might fit into several categories. In these situations, the best possible fit was chosen. The process was highly opinion-based, and the evaluation of one individual might impair the study's quality and liability. The classification schema was presented and evaluated by two viewers to ensure the research quality, but the classification was performed alone.

4.2 Results of Classification

When the classification scheme is established, the actual data extraction takes place, and the articles are sorted into different classes (Petersen et al., 2008). Typically, the classification schema evolves during the data extraction when the papers are further exanimated (Petersen et al., 2008). In this paper, the main framework was kept the same through the classification. However, some of the class names and definitions were modified, and the additional layer of impact (presented in Table 6) was added to the classification schema. The classification was performed in a spreadsheet with comments and notes from previous paper reviews. The cleaned version of the spreadsheet is presented in Table 7.

In the classification, a significant portion of papers focused on biased algorithms. These papers were classified in pertinence facet to "full" if the papers focused on making the whole system more transparent. Papers that focus on cleaning and fixing the biased datasets are classified as having a "partial" pertinence towards XAI. They were considered to have the main focus more on data science. The pertinence facet helped to understand if the paper has a strong focus on XAI and transparency issues. Around half of the papers (53%) had full focus on XAI. Papers with a marginal focus to 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 in different classes were calculated. The classification results are presented in Table 6 with the number of papers in each facet's class and the percentage of the class compared to the full sample (n=76). This method visualizes what has been emphasized in past research, the research gaps, and the possibilities for future research (Petersen et al., 2008).

TABLE 6 Results of Classification

Results of Classification		
Research facet		
Proposal	46	61%
Philosophical	17	22%
Experience	10	13%
Validation	3	4%
Contribution facet		
Model	26	34%
Tool	29	38%
Procedure	12	16%
Advice	8	11%
Specific solution	1	1%
Focus facet		
Bias	37	49%
Black Box	18	24%
Accountability	2	3%
Attitudes	18	24%
Pertinence facet		
Full	40	53%

Partial	26	34%
Marginal	10	13%

In the research facet, the proposal class was significantly emphasized, with 61% of the studies proposing a technical, mathematical, or design solution. The number of philosophical papers structuring and framing the problem area was the second-largest class. Experience papers and validation were less frequent, which indicates the immaturity of the research field, as many issues are not yet implemented or studied in practice.

The main contribution classes were models (structuring the problem area) and tools (computational solutions to a particular problem). Many of the papers proposing a new computational tool were proposing a new algorithm or mathematical solution. Also, procedures (proposals of design methods and frameworks) and advices (more directive proposals than ready solutions) were visible in the sample. Solutions to specific application problems were proposed only in two papers, which could again indicate the study field's immaturity.

Papers focused mainly on bias algorithms (49%) and black boxes and other non-transparent systems (24%). Papers where the main focus was to understand developers' and users' expectations, attitudes, and trust towards the explainable AI systems represented 24% of the whole sample. There were two more papers studying users, but those papers' main focus was in black-box models; hence, they were categorized into black-box. From the Attitudes category, only six papers (8% of the whole sample) focused on practitioners' expectations and opinions, and the rest 12 papers focused on understanding how the general public is seeing the issue. Accountability was the main focus only in two papers but was a secondary focus in several papers. The results are visualized in Figure 7, which shows the number of papers in each category. More indebted visualizations and analyzations are provided in the following chapter.

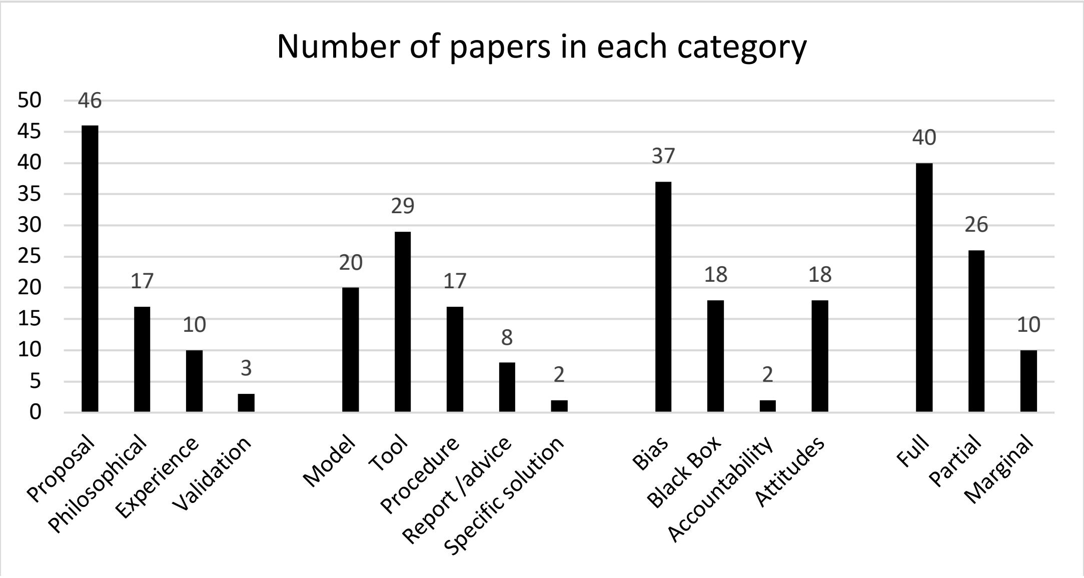


FIGURE 7 Visualization of Results of Classification

Together with the classification, the paper's social impact was studied with two filters; the type of data used (real-world data vs. synthetic data) and if the paper paid attention to societal issues or the interest was purely in technical challenges. The impact classification results are presented in Table 7, where the number refers to the number of papers in each row and the percentage to the portion of the whole sample (n=76).

TABLE 7 Research of Connection to Real-World Issues

Results of Classification		
Impact		
Use of real-world data	67	88%
Contributing societal issues	66	87%

Sixty-seven papers used either real-world datasets or collected the data in their research. That is the majority of papers, 88%. Societal issues were in the center of attention in 66 papers, 87% of the whole sample (n=76%). The strong emphasis on societal issues is not surprising because the papers were selected to the sample if they focus on AI's ethical issues. The amount of real data used is an interesting factor, as it indicates the close connection to real-world problems.

4.3 Overview of final sample

The overview of the final sample (n=76) with the classification results is presented in Table 8. The only document identifier shown is the first author and the year of publication to simplify the visualization. The Classification columns show the primary category in each classification facet. The Impact columns show the usage of data used (real vs. synthetic) and if the paper takes a stand on societal issues (yes vs. no). All the papers are found in the reference list at the end of this paper.

TABLE 8 Classified Dataset

Paper	Classification				Impact	
1st Author	**Research**	**Contribution**	**Focus**	**Pertinence**	**Data**	**Societal**
Caliskan et al (2017)	Proposal	Model	Bias	Partial	Real	Yes
Babu et al (2018)	Proposal	Tool	Bias	Partial	Real	Yes
Calmon et al (2018)	Proposal	Tool	Bias	Partial	Real	Yes

Dixon et al (2018)	Proposal	Tool	Bias	Full	Real	No
Ehsan et al (2018)	Proposal	Tool	Black box	Full	Real	No
Flexer et al (2018)	Validation	Model	Bias	Full	Real	Yes
Grgić-Hlača et al (2018)	Proposal	Procedure	Bias	Full	Real	Yes
Grgic-Hlacaet al (2018)	Experience	Model	Attitudes (public)	Partial	Real	Yes
Henderson et al (2018)	Philosophical	Model	Bias	Partial	Real	Yes
Iyer et al (2018)	Proposal	Tool	Black box	Full	Real	Yes
Raff et al (2018)	Proposal	Tool	Bias	Full	Real	Yes
Shank et al (2018)	Philosophical	Model	Attitudes (public)	Marginal	Real	Yes
Srivastava et al (2018)	Proposal	Procedure	Bias	Full	Real	Yes
Veale et al (2018)	Experience	Model	Attitudes (developers)	Full	Real	Yes
Yang et al (2018)	Proposal	Tool	Bias	Full	Real	Yes
Zhang et al (2018)	Proposal	Tool	Bias	Full	Real	No
Zhou et al (2018)	Philosophical	Model	Attitudes (public)	Marginal	Real	Yes
Abeywickrama et al (2019)	Proposal	Procedure	Accountability	Partial	Real	Yes
Addis et al (2019)	Experience	Advice	Attitudes (developers)	Full	Real	Yes
Aïvodji et al (2019)	Proposal	Tool	Black box	Full	Real	Yes
Ali et al (2019)	Proposal	Tool	Bias	Full	Real	Yes
Amini et al (2019)	Proposal	Tool	Bias	Partial	Real	Yes
Barn (2019)	Philosophical	Model	Attitudes (public)	Marginal	Real	Yes
Beutel et al (2019)	Proposal	Tool	Bias	Partial	Real	No
Bremner et al (2019)	Proposal	Tool	Black box	Partial	Real	Yes
Brunk et al (2019)	Proposal	Model	Black box	Full	Real	Yes
Cardoso et al (2019)	Proposal	Tool	Bias	Full	Synthetic	Yes
Celis et al (2019)	Validation	Model	Bias	Partial	Real	Yes
Coston et al (2019)	Proposal	Tool	Bias	Full	Real	Yes
Crockett et al (2019)	Philosophical	Model	Attitudes (public)	Partial	Real	Yes
Garg et al (2019)	Proposal	Tool	Bias	Partial	Real	Yes
Goel et al (2019)	Proposal	Tool	Bias	Partial	Real	Yes
Green et al (2019)	Philosophical	Advice	Attitudes (public)	Marginal	Real	Yes
Heidari et al (2019)	Philosophical	Advice	Bias	Partial	Real	Yes
Hind et al (2019)	Proposal	Procedure	Black box	Full	Synthetic	No
Kim et al (2019)	Proposal	Tool	Black box	Full	Real	Yes
Lai et al (2019)	Philosophical	Model	Attitudes (public)	Partial	Real	Yes
Lakkaraju et al (2019)	Proposal	Procedure	Black box	Full	Real	Yes

Lux et al (2019)	Proposal	Tool	Bias	Full	Real	Yes
Mitchell et al (2019)	Proposal	Procedure	Bias	Full	Synthetic	Yes
Noriega-Campero et al (2019)	Proposal	Tool	Bias	Full	Real	Yes
Radovanović et al (2019)	Proposal	Specific solution	Bias	Partial	Real	Yes
Raji et al (2019)	Validation	Tool	Bias	Full	Real	Yes
Rubel et al (2019)	Philosophical	Model	Accountability	Full	Real	Yes
Saxena et al (2019)	Philosophical	Advice	Attitudes (public)	Marginal	Real	Yes
Sivill (2019)	Philosophical	Advice	Bias / av	Partial	Real	Yes
Srinivasan et al (2019)	Proposal	Tool	Bias	Partial	Real	Yes
Teso et al (2019)	Proposal	Procedure	Black box	Full	Real	Yes
Ustun et al (2019)	Proposal	Tool	Bias	Partial	Real	Yes
Vakkuri et al (2019)	Experience	Procedure	Attitudes (developers)	Marginal	Real	Yes
Vanderelst et al (2019)	Philosophical	Advice	Attitudes (public)	Marginal	Real	Yes
Vetrò et al (2019)	Philosophical	Advice	Bias	Partial	Real	Yes
Webb et al (2019)	Philosophical	Model	Attitudes (public)	Full	Real	Yes
Wolf et al (2019)	Proposal	Model	Black box	Full	Synthetic	No
Wouters et al (2019)	Experience	Model	Attitudes (public)	Partial	Real	Yes
Yilmaz et al (2019)	Proposal	Tool	Black box	Full	Synthetic	Yes
Balachander et al (2020)	Proposal	Specific solution	Black box	Full	Real	Yes
Brandão et al (2020)	Proposal	Procedure	Bias	Full	Real	Yes
Chen et al (2020)	Proposal	Tool	Bias	Full	Synthetic	No
Clavell et al (2020)	Experience	Tool	Bias	Full	Real	Yes
Cortés et al (2020)	Proposal	Procedure	Bias	Full	Synthetic	Yes
He et al (2020)	Proposal	Tool	Bias	Full	Real	No
Jo et al (2020)	Experience	Procedure	Bias	Marginal	Real	Yes
Karpati et al (2020)	Philosophical	Advice	Black box	Full	Real	Yes
Lakkaraju et al (2020)	Proposal	Procedure	Black box	Full	Real	Yes
Madaio et al (2020)	Experience	Model	Attitudes (developers)	Marginal	Real	Yes
Mitchell et al (2020)	Proposal	Procedure	Bias	Partial	Real	Yes
Nirav et al (2020)	Philosophical	Procedure	Attitudes (public)	Marginal	Real	Yes
Orr et al (2020)	Experience	Model	Attitudes (developers)	Partial	Real	Yes
Schelter et al (2020)	Proposal	Procedure	Bias	Partial	Real	Yes
Sendak et al (2020)	Philosophical	Model	Black box	Full	Real	Yes
Sharma et al (2020a)	Proposal	Procedure	Black box	Full	Synthetic	No
Sharma et al (2020b)	Proposal	Tool	Bias	Partial	Real	Yes

Shulman et al (2020)	Proposal	Tool	Black box	Full	Synthetic	No
Slack et al (2020)	Proposal	Procedure	Black box	Full	Real	Yes
Vakkuri et al (2020)	Experience	Model	Attitudes (developers)	Partial	Real	Yes

In the following chapter, the classified data is analyzed and visualized. The analysis aims to better understand the study field of XAI and its role in AI ethics research. Before moving to analyzes, one more perspective from the classification is presented.

4.4 Explainability vs Interpretability in Black Box Papers

As described in this paper's background section, there are two approaches to ensure the transparency in black box systems; explainability (provide an understandable explanation) and interpretability (transparency of the whole decision process). A short analysis was performed to understand how the topic was interpreted in the sample of this paper. The evaluation was performed separately from the classification as it was only performed to papers with the main focus in black box because the rest of the papers did not take a stand on the issue. The papers focusing on Black box issues were evaluated to see if they focus more on explainability or interpretability. This categorization does not take a stand on whether the paper is providing a solution or pointing out challenges in the area of focus or not. Table 9 presents the results of the categorization.

TABLE 9 Perspective in Black Box Papers

Perspective in Black Box papers	
1st Author	**Perspective**
Wolf et al (2019)	explainability
Hind et al (2019)	explainability
Teso et al (2019)	explainability
Aïvodji et al (2019)	explainability
Sharma et al (2020)	explainability
Slack et al (2020)	explainability
Shulman et al (2020)	explainability
Iyer et al (2018)	interpretability
Lakkaraju et al (2019)	interpretability
Yilmaz et al (2019)	interpretability
Kim et al (2019)	interpretability
Bremner et al (2019)	interpretability

Sendak et al (2020)	interpretability
Balachander et al (2020)	interpretability
Karpati et al (2020)	interpretability
Brunk et al (2019)	user study
Lakkaraju et al (2020)	user study

Two of the papers researched the effect of transparency on the users and did not take a stand on how the explanations were generated. These are categorized as user studies in this section. From the remaining 15 papers, seven papers focused on explainability and eight papers to interpretability. The result indicates that when the focus of AI ethics is present in the research, and the research uses empiric material, both approaches (explainability and interpretability) are equally present in the corpus.

> EC3: The Black Box problem is researched equivalently from the perspectives of interpretability and explainability.

No further analyzes were performed on this sub-part of classification. The analyzes of full classification are presented next.

5 SYSTEMIC MAP

There are several ways to visualize the results of systemic mapping study. The two most common are bar plots and bubble plots (Petersen et al., 2015). To illustrate the number of studies for a combination of categorizations, bubble plots visualization is exceptionally well suited (Petersen et al., 2015). 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. The SMS is not tied in bubble and bar plots, and further visualization alternatives can be found in statistics, Human-Computer Interaction field, and information visualization fields (Petersen et al., 2008). As there were four main facets in the classification schema, it was required 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 chapters, the results of classification schema, pertinence, impact, annual change, and the venue of the study field are visualized and analyzed.

5.1 Systemic Map in the Bubble Plot Visualization

A bubble plot diagram helps to give a quick overview of the research field and support the analyzes better than the frequency tables. The bubble plot diagram is built by using summary statistics from Table 5 presented in the previous chapter. The diagram visualizes the frequencies and correlations between categories and facets. The bubble plot diagram is two x-y scatterplots with bubbles in category intersections. The same idea is used twice, on the opposite sides of the same diagram, to show the intersection with the third facet on the x-axis. In this case, contribution and research facets are compared to the focus facet. The size of a bubble indicates the number of papers that are in the intersection of the coordinates. Next to a bubble, there is the percentage of the total amount (n=76) in the represented category of the x-axis. The bubble plot is presented in Figure 8. (Petersen et al., 2008).

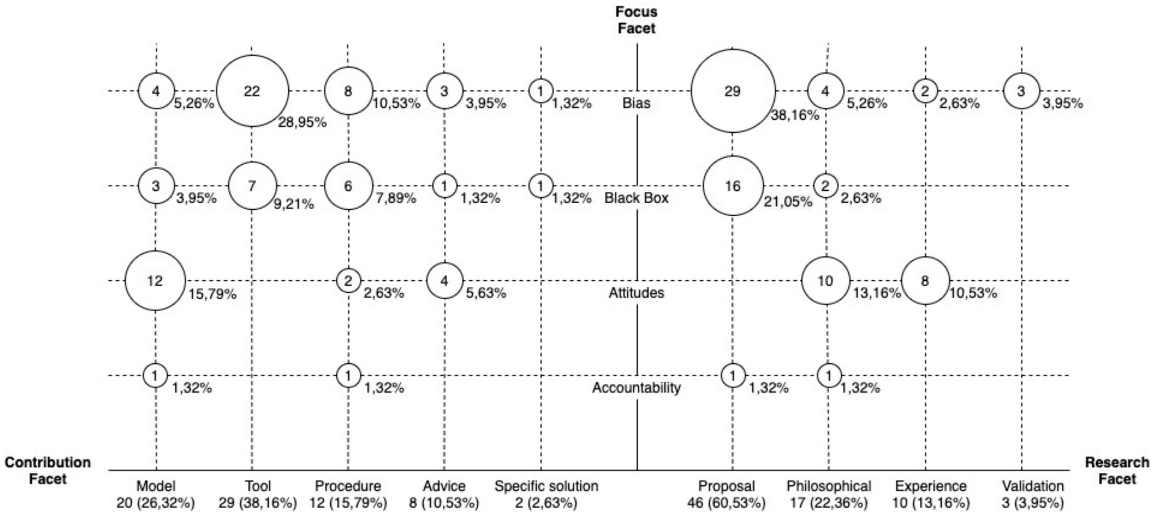


FIGURE 8 Visualization of Systemic Map in the Form of Bubble Plot

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

> EC4: The most popular paper type in the research facet is a proposal for solving algorithmic bias.

Also, the proposals for black-box issues are highlighted. Proposal researches study new, novel techniques to solve a particular issue. If compared with the validation research, which studies a specific solution that has already been implemented in practice, the size of the proposals bubble is much larger, which again points to the research field's freshness. Either there are no proper practical solutions to fix the ethical issues related to XAI, or these solutions are not yet implemented in practice, or the practical implementation is not yet studied. Probably the reason for scarcity in validation research is partly due to all the reasons mentioned above.

From the Contribution facet, the largest bubble can be found in the intersection of bias and tool. Little less than a third (22 papers) of the whole sample contributes to the research field with a computational solution to solve algorithmic biases.

> EC5: Almost a third of the papers in the whole sample contribute to the research field with a computational solution to solve algorithmic biases.

A computational tool to solve black-box issues has been proposed in 7 papers. In the intersection of the attitude facet and the model facet is the second-largest bubble with 12 papers. The bubble visualizes how the research field is modeled and structured by providing a better understanding of users and practitioners. Procedures, contributing with a new way to solve issues such as design frameworks, have been quite frequent in the focus areas of bias algorithms and black-box problems.

PEC2: In the study field of Explainable Ethical AI, the most common type of empiric research is to study a novel technique that can solve a computational challenge.

There is no precise weighting on any of the contribution types in the black box's focus category. Compared to the Bias category with apparent weight in the contribution of computational tools and attitudes category with an apparent weight in modeling the problem area. From the focus category of bias with 37 papers (49% of the whole sample), 20 papers (26% of the whole sample) are in the research category proposal and contribute with a computational tool. That is the most distinctive type of papers in this research.

EC6: The most distinctive paper type is a computational tool proposing a solution to a problem with bias.

As a conclusion of the bubble plot, the most common type of paper is a computational tool proposing to solve a problem with biases, and in general, the majority of the papers look for novel techniques and solutions to computational problems. The results might indicate that the focus is slightly monotonous. Papers concerning black boxes, accountability, or attitudes are more distinctive except the strong emphasis on proposals as a research type in the black box category. Also, the results indicate the immaturity in the research field.

EC7: The research field seems a bit monotonous and immature in considering the variety of topics, used research methods and contributions of the papers.

5.2 Pertinence Mapped in Par Plot

As the pertinence indicates the accuracy with XAI's topic, the pertinence was visualized with a bar plot corresponding to the focus facet. This visualization aims to understand in which focus areas the research field has full pertinence on XAI and transparency-related topics, and in which focus areas the pertinence is elsewhere. The bar plot is in Figure 9. The size of the plot presents the number of papers from the full sample (n=76) in each category.

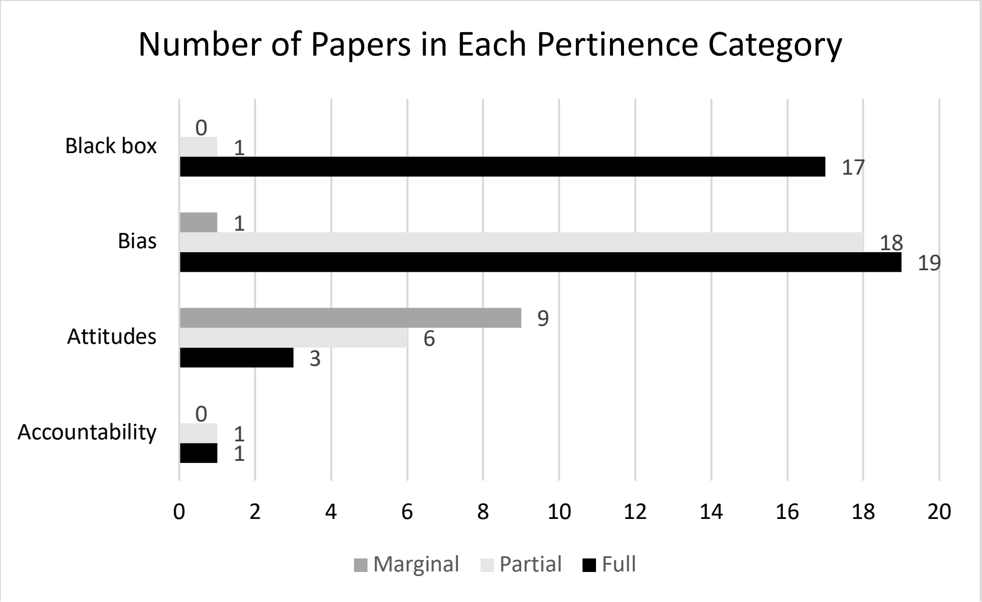


FIGURE 9 Pertinence of Focus Areas

Not surprisingly, most of the papers (17 out of 18 papers) focusing on the black box were categorized to have a full focus on XAI. The black box is one of the core concepts in XAI research (Adadi & Berrada, 2018).

> EC8: From the papers focusing on black box (n=18) 94,5% had full pertinence on XAI.

From the papers focusing on algorithmic biases, 19 had full focus, 18 partial focus, and one marginal focus to XAI. Many of the papers with partial focus had the main focus on cleaning data and fixing the datasets that are causing the discriminative and unfair decisions. These papers were considered to have the main pertinence in data science and fairness rather than in XAI. Accountability was the main focus in only two papers, so no interpretation is possible to make.

Interestingly only three papers focused on the attitudes and expectations of practitioners, users, and the public had full pertinence towards XAI. It is important to note that two more papers were categorized with black box focus, but that had a strong focus on understanding people's perceptions and attitudes. With Attitudes as the main focus, six papers had partial and nine papers marginal relation to XAI. The results indicated a research gap in understanding people's perceptions of the topic.

PEC3: The human perspective towards Explainable AI is not well-known. There is a lack of research about the practitioners' and user's expectations and attitudes towards Explainable AI.

Compared to autonomous vehicles, another widely researched ethical issue of AI, the research field of XAI differs. There are several profound studies in the autonomous vehicles field, such as Moral Machine (Awad et al., 2018), aiming to understand human morality and expectations towards autonomous systems. The research field of XAI seems to be a bit distant from users' expectations. Especially the attitudes and expectations of practitioners and developers have not been profoundly studied. Only two papers with the main focus on the expectations of practitioners had a full pertinence to XAI. Based on these findings, it is assumed that in the research of XAI in AI ethics, there is a lack of understanding of the issues related to practical implementation and practitioners' attitudes. Only one paper studied the current state of industrial implementation of AI ethics in general, and none with full pertinence to XAI. No paper studied the managerial or business perspective of XAI.

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

In conclusion, the pertinence is strongest in black box research, strongly present in the bias category, and in the accountability category, which is so small that any interpretations cannot be drawn. The attitudes category had a relatively weak connection to XAI. This indicates a need to understand better how people, both the practitioners, businesses, and the public, perceive XAI.

5.3 Analysis of Synthetic Data Use and Societal Perspective

To understand if the study field focuses on real-world problems, the papers were evaluated based on if they were taking a stand on societal issues, and if the data used was real-world data or synthetic data. As mentioned at the end of chapter four, only 12% (9 papers) used synthetic data. All the papers using synthetic data had a proposal as the research type, and all had full pertinence to XAI. The focus varied between Black Box (5 papers) and Bias (4 papers), and the contribution of the papers was a model (1 paper), procedure (4 papers), or tool (4 papers). Four papers took a stand on societal issues, and five papers were purely technical. The results are visualized in Figure 10. The number in each section refers to the number of papers from the sample of synthetic data papers (n=9).

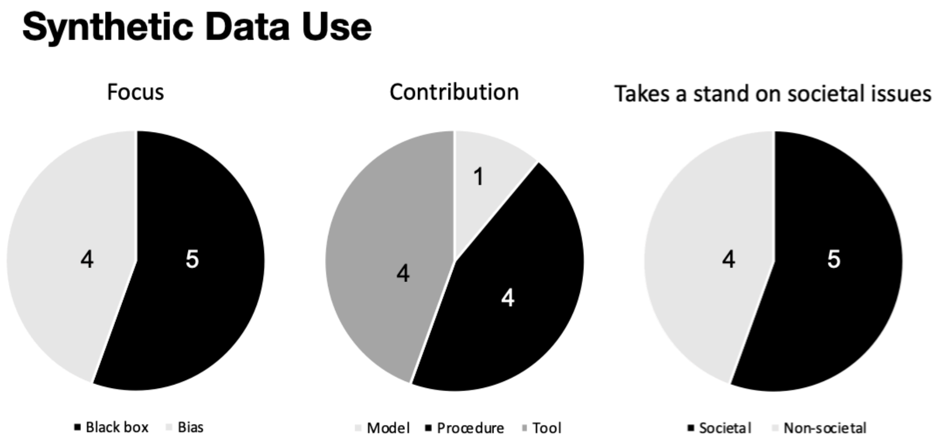


FIGURE 10 Number of Papers Using Synthetic Data (n=9)

Interestingly, 27,7% of papers focusing on the black box (18 papers in total focused on the black box) used synthetic data. Compared to papers focusing on bias where the percentage of papers using synthetic data is only 10,8% (5 papers out of a total of 37 papers focused on the bias).

> EC9: Papers focusing on the black boxes more often use synthetic data than papers that focus on biases.

This could indicate that the black box problem is researched more as a technical challenge, and algorithmic bias research closer to the real-world. Alternatively, there might be a wider variety of real-world datasets available to study algorithmic fairness.

The societal perspective had a strong presence in the research field. Only 13%, ten papers, did not take a stand on societal issues. Similarly, as with synthetic data use, all of these papers had the proposal as a research type. Nine out of ten papers had full pertinence to XAI. 50% (5 papers) had a focus on Biases and 50% in Black Box. In seven papers, the contribution was a tool. In two papers, it was the procedure, and in one a model. The tool category's emphasis could indicate that papers providing a computational solution more rarely take a stand on societal issues. From a total of 29 papers with a contribution to a tool, this is 24,1%. Wherein the two papers contributing to a procedure, that is only 11,8% of the total of procedure papers (17 papers). In five papers, real data was used, and in five papers, synthetic. The results are visualized in Figure 11. The number in each section refers to the number of papers from the sample of no societal contribution (n=10).

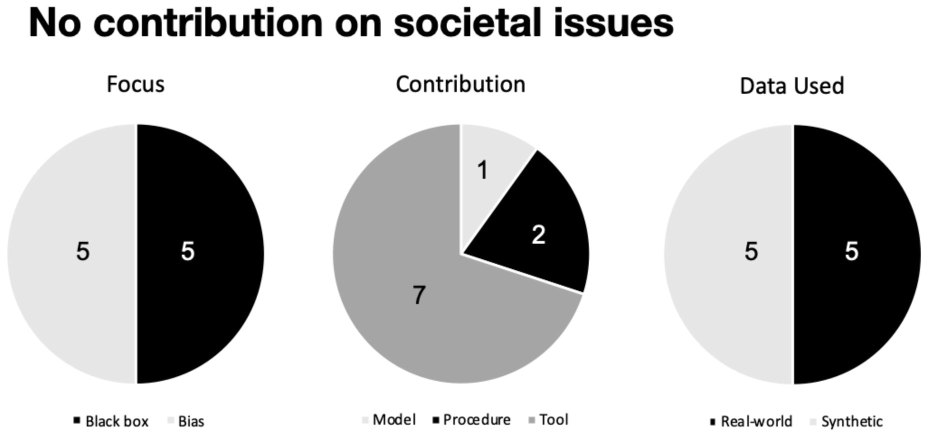


FIGURE 11 Number of Papers with No Contribution on Societal Issues (N=10)

Notably, only one paper focusing on biases without societal contribution used synthetic data. That is 2,7% of total papers in the bias category (n=37). That indicates that papers with the focus on biases were closely connected to real-world problems, either by using real-world data or by contributing to real-world societal issues. Similarly, papers focusing on attitudes and accountability only used real-world data and always took a stand on societal issues. Four papers focusing on the black box used synthetic data and did not take a stand on societal issues. That represents 22.2% of the papers focusing on the black box (n=18).

> EC10: Around a fifth, (22,2%) of the papers focusing on black boxes used synthetic data and were not contributing to societal issues, hence had a distant connection to the real-world problems.

Overall, the research field is close to real-world problems and interested to contribute to society.

> PEC5: Explainable AI researchers are interested in real-world applications, not only technical aspects of the topic. The empiric research area of Explainable AI has a close connection to real-world problems when the research is related to AI ethics.

In conclusion, there is a strong connection to real-world and societal issues in the research area. The only area with a slight indication of distancing from the real-world issues was the research area focusing on the black box. If XAI's research area were studied independently without the association with AI ethics, the connection to real-world problems might have been different.

5.4 Visualization of Annual Changes in the Research Field

The year range for the SMS performed in this paper was 2012-2020, but none of the papers from 2012-2016 were included in the study after inclusion and exclusion. From 2017 only one paper was included, 2018 16 papers, 2019 39 papers, and from the first half of the year 2020 20 papers. The annual changes are visualized in Figure 12, where the plot's size presents the number of papers published per year (n=76). The results indicate the growth of the research field and XAI's freshness in the AI ethics research field with empirical results.

> EC11: Explainable AI is young but growing empiric research area in the field of AI ethics.

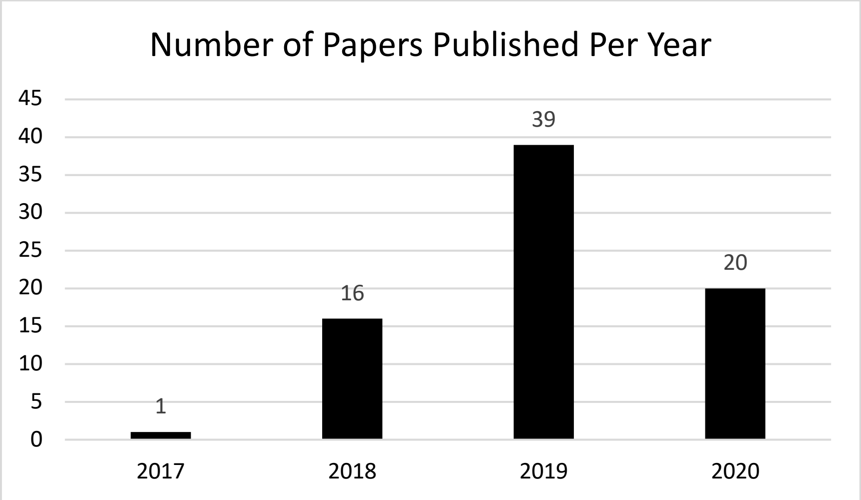


FIGURE 12 Yearly distribution of included papers

To visualize further the annual changes in the research field, two Bubble plots were created. Figure 13 shows the annual changes and evolution in the contribution and research facets. Figure 14 shows the annual changes in the focus and pertinence facets. The plot's number refers to the number of papers in the intersection and the percentage next to the plot compared to the full sample (n=76). The motivation for bubble plots was to detect trends in the research field. Although, as the research field is young, the trends might be only seasonal changes. Moreover, because the year 2020 cannot be evaluated entirely, the results per year are not fully comparable.

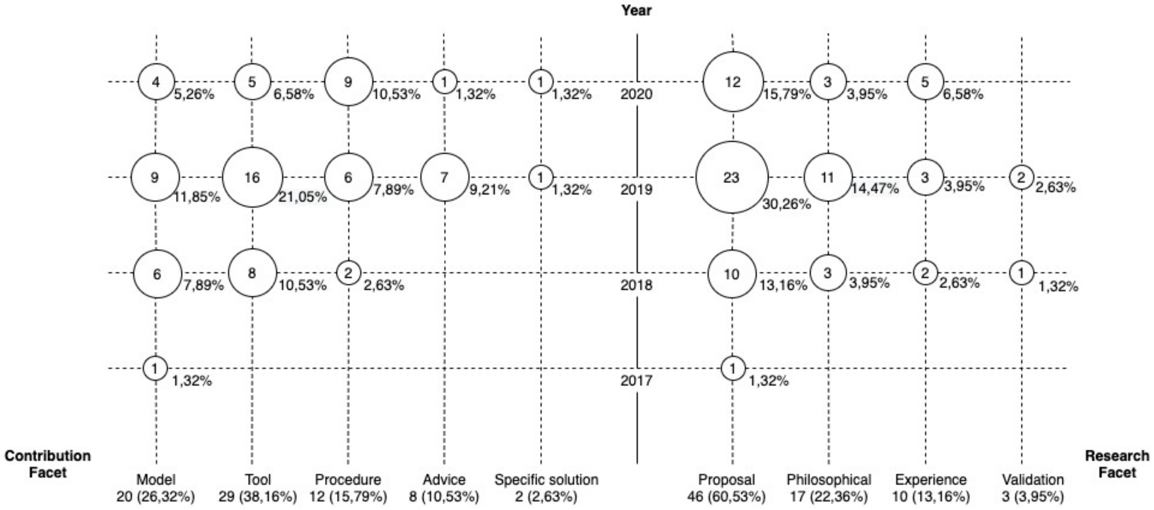


FIGURE 13 Annual Changes in the Research and Contribution Facets

The bubble plot reveals that the proposal has been the most popular category from the research facet every year. Experience papers seem to be getting more popular as the research field matures, and in 2020, more experience papers have been published than philosophical papers. The research trend seems to be towards more practical understanding and less philosophical framing and structuring of the focus area.

> EC12: The trend is towards 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 of better ways to do something. Interestingly the tools, the computational solutions, seem to have been decreasing in 2020. That could indicate that the research field is evolving to become more holistic and not as intensely focused on finding technical solutions. However, more research is required to verify the conclusion. Another exciting factor is that advice papers seem to be decreasing even though the research field is maturing.

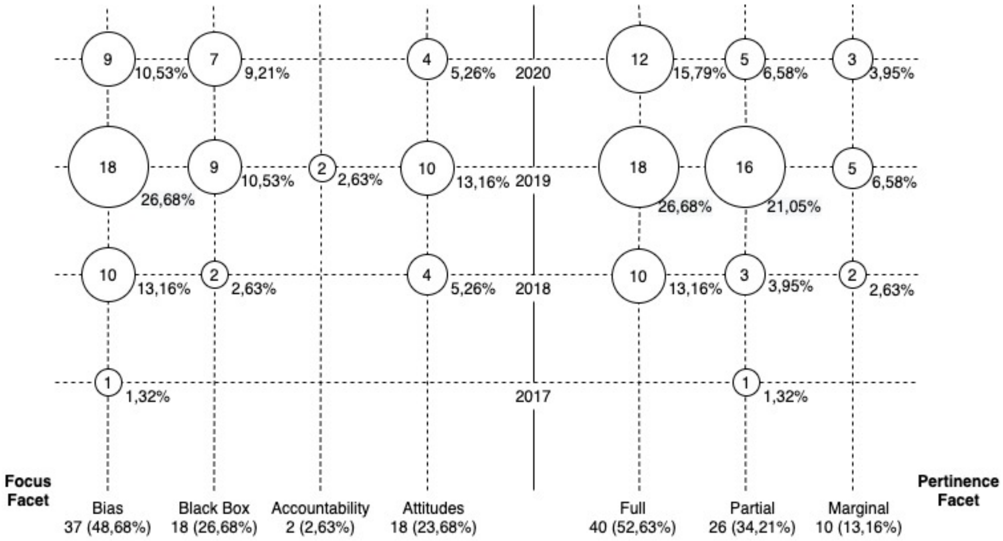


FIGURE 14 Annual Changes in Focus and Pertinence Facets

The bubble plot visualizing annual changes in pertinence and focus facets shows a slight trend towards Full pertinence in XAI in 2020. That could indicate that AI ethics research focusing on XAI is also growing independently, not only because of the growth in related fields, such as the research of avoiding biases by detecting flaws in the dataset. However, the sample is too small, and the research field too young to confirm that the trend is not just a seasonal variation.

The focus category bias has doubled between 2018 and 2019, and if the pace of publication stays the same, it is assumed to double in 2020. Black box research seems to accelerate, as in the first six months of 2020, already seven papers are published, which is almost as much as in 2019 when nine papers with black box focus were published. That could indicate the black box research to grow almost as large as bias research in 2020. The results imply that black box research is the fastest growing trend in the research area of XAI.

> EC13: Black box research seems to be the fastest-growing focus area in the research field of Explainable AI.

The number of papers focusing on attitudes seems to stay relatively similar in 2020 as in 2019, with no expected growth. From the attitude papers, the annual division of papers focusing on understanding the developers and practitioners is: 1 paper in 2018, 2 papers in 2019, and 4 papers in 2020. Understanding the expectations, needs, and opinions of practitioners seems to be a growing trend. That could indicate that the research field is more and more interested in practical implementation.

> EC14: There is a growing interest in practical implementation and understanding of the needs and expectations of users and practitioners.

In conclusion, the amount of papers is growing annually. Especially the proposal research type is getting more popular, and the popularity of papers with the contribution of a procedure or a tool is growing. Papers focusing on the black box seem to reach the popularity of focus on the bias. More papers with full focus on XAI every year indicate that the research area is shifting from a secondary focus area to become a primary focus area of research.

EC15: In the field of AI ethics, the research area with a primary interest in Explainable AI is growing.

5.5 Venue 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. Journals are typically the premier publication venue in software engineering; thus, the papers published in journals should include the most mature research (Ivarsson & Gorschek, 2010). Also, a higher degree of empirical evidence is expected from papers published in journals than from the conference of workshop proceedings (Ivarsson & Gorschek, 2010).

The majority of the papers, 59 out of 76 (77,3%), were conference proceedings. The rest 17 papers (22,4%) were published in journals. The most popular venue was AAAI/ACM Conference on AI, Ethics, and Society (AIES). Nearly half, 43,4%, 33 papers, of the total sample (n=76), was published in AIES.

PEC6: In the corpus of explainable ethical AI, the publication venue of empirical research is monotonic, with 43,4% of the papers published in one conference.

The second most popular venue was the Conference on Fairness, Accountability, and Transparency (FAT*) with five paper publications. There were no other conferences with more than one or two proceedings from the sample. The journal papers were all published in different journals except for two papers published in Emerald's Journal of Information Communication & Ethics in Society, but in different volumes.

The line chart in Figure 15 visualizes the annual growth in the publication venue. The black line represents the number of papers published each year in conferences (n=59), and the grey line shows the papers published in journals (n=17). The growth in conference proceedings has been slightly faster than the growth in journal publications, but the comparison between journals and conference proceedings seems regular.

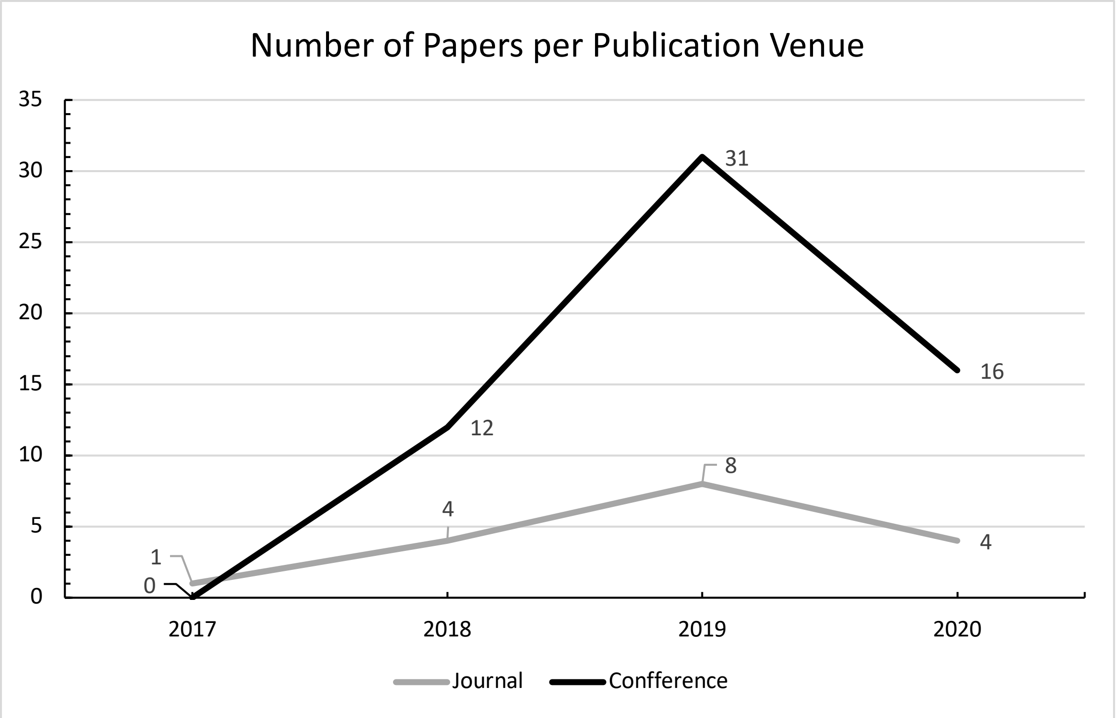


FIGURE 15 Annual Changes in Publication Venue

Interestingly, out of 17 papers published in journals, eight focused on attitudes. That is a large proportion of Attitude papers; 44,4% to be exact (n=18). As the rigor in journal publications is higher (Ivarsson & Gorschek, 2010), this indicates that even though the field lacks in a plurality of studies in human's role and attitudes, the quality of that type of research is high.

EC16: The studies of the role of humans are rare but high-quality research.

This reflection might be explained with the type of data used in the research. User research usually requires a more time-consuming research method; hence the originality and quality of the evidence are higher, which fits better with the publication criteria of journals. Compared to the black box papers where only 11,1% (2 papers) were published in journals and the bias papers where 16,2% (6 papers) were published in journals. From the focus category of accountability, one paper was published in the journal, and one was a conference proceeding.

In conclusion, the publication venue's reflection reveals a monotony in the publication, as nearly half of the papers were published in the same venue. The ratio between journal publications and conference proceedings seems to be aligned. Nevertheless, the focus areas were not equally presented in journals, where the strong emphasis was on attitude studies, wherein from black box papers, only 11,1% were published in journals.

EC17: Black Box research in the AI ethics field is rarely published in journals.

5.6 Summary of empirical conclusions

This chapter summarizes the empirical conclusions and primary empirical conclusions of this paper. This paper's main theoretical contribution is mapping the research area, which supports future research by framing and visualizing the existing research. The secondary contribution is the primary empirical conclusions (PEC) derived from the maps. Primary empirical conclusions are supplemented with empirical conclusions (EC). Empirical conclusions that were highlighted from the text body in pervious chapters are listed in Table 9.

TABLE 10 List of Empirical Conclusions

List of Empirical Conclusions	
Identifier	**Empirical Conclusion**
EC1	Most of the research papers in the field of AI ethics do not use empiric evidence. Only 20% of the papers provide empirical evidence.
EC2	Empiric Research of AI ethics grew significantly in 2018, following the trend in public discussion.
EC3	The Black Box problem is researched equivalently from the perspectives of interpretability and explainability.
EC4	The most popular paper type in the research facet is a proposal for solving algorithmic bias.
EC5	Almost a third of the papers in the whole sample contribute to the research field with a computational solution to solve algorithmic biases.
EC6	The most distinctive paper type is a computational tool proposing a solution to a problem with bias.
EC7	The research field seems a bit monotonous and immature when considering the variety of topics, used research methods, and contributions of the papers.
EC8	From the papers focusing on the black box (n=18), 94,5% had full pertinence on XAI
EC9	Papers focusing on the black boxes more often use synthetic data than papers that focus on biases
EC10	Around a fifth, (22,2%) of the papers focusing on black boxes used synthetic data and were not contributing to societal issues, hence had a distant connection to the real-world problems.
EC11	Explainable AI is a young but growing empiric research area in the field of AI ethics
EC12	The trend is towards more practical implications and less philosophical framing of the focus area
EC13	Black box research seems to be the fastest-growing focus area in the research field of Explainable AI.
EC14	There is a growing interest in practical implementation and understanding of the needs and expectations of users and practitioners.
EC15	In the field of AI ethics, the research area with a primary interest in Explainable AI is growing
EC16	The studies of the role of humans are rare but high-quality research.

EC17	Black Box research in the AI ethics field is rarely published in journals.

The primary empirical conclusions are listed in Table 10. In previous chapters, the primary empirical conclusions were boxed from the text body to bring them into the reader's attention and ensure easy findability when skimming the paper. Primary empirical conclusions are written in a context-enriched manner to support the understanding of the readers that are not familiar with the full paper.

TABLE 11 List of Primary Empirical Conclusions

List of Primary Empirical Conclusions	
Identifier	**Primary Empirical Conclusion**
PEC1	Explainable AI is significant research focus on the study field of AI Ethics. From the empiric research papers published after 2012, 36.2% is related to Explainable AI.
PEC2	In the study field of Explainable Ethical AI, the most common type of empiric research is to study a novel technique that can solve a computational challenge.
PEC3	The human perspective towards Explainable AI is not well-known. There is a lack of research about the practitioners' and user's expectations and attitudes towards Explainable AI.
PEC4	Industrial implementation of Explainable AI is not profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of Explainable AI.
PEC5	Explainable AI researchers are interested in real-world applications, not only technical aspects of the topic. The empiric research area of Explainable AI has a close connection to real-world problems when the research is related to AI ethics.
PEC6	In the corpus of explainable ethical AI, the publication venue of empirical research is monotonic, with 43,4% of the papers published in one conference.

Theoretical and practical implications of the primary empirical conclusions are evaluated next in Discussion.

6 DISCUSSION

This chapter lists the proposals for the theoretical and practical implications for the Primary Empirical Conclusions (PEC), which were the SMS process outcomes. In theoretical implications, PEC's are mirrored to the existing research. The practical implications are proposals and ideas, how these conclusions could be implemented in practice.

6.1 Theoretical Implication

The main theoretical implication of this paper is the mapping of the research area presented in chapter 6. The key outcomes of the analysis of the mapping process are in this chapter mirrored with existing research. Primary empirical conclusions 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 paper is to understand the research area's scope and depth, rather than the quality of the articles, the primary empirical conclusions are related to those factors.

The significant proportion of papers related to XAI in the empiric research of AI ethics (PEC1) corresponds to the research of Jobin et al. (2019), which noted that the transparency is the most frequently highlighted principle in AI ethics. Besides, the result reflects the overall importance and interest of XAI. Simultaneously it reflects XAI's connection to real-world problems as it is studied with empirical methods.

As far as the author knows, there is no previous research that analyzes the type of research done in the field yet, so the relation to existing research might be shallow. 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 is interested in finding technical solutions to ethical problems (Brundage, 2014), which correlates with a broader perspective.

Human's role and perspective are understudied subjects, both the user and the practitioners' point of view (Ferreira and Monteiro 2020; Adadi and

Berrada, 2018). The same finding was done 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. (2020) pointed out the same dilemma with AI ethics. Their research is the only paper in this SMS with current state research in the practical implementation of ethical principles.

Unlike black box problems, where the research field is distancing from real-world problems (Rudin, 2019), the XAI has a strong contribution to real-world problems (PEC5). The slight distancing in black box research was also reflected in the results of this research. Still, the vast majority of the papers focusing on the black boxes were contributing to societal issues and/or using real-world data in their research.

SMS's somewhat surprising reflection was the monotony in the research venue of empiric research (PEC6). The research field of AI is interdisciplinary, similar to that of AI ethics (Russell & Norvig, 1994; Vakkuri & Abrahamsson, 2018). The fragmentation of the field even makes studying the field rather challenging. Therefore, it was not expected that almost half of the papers in empiric research of ethical XAI would be published in the same conference, AIES, even though the conference is cross-disciplinary. The result does not imply that the AIES is not a suitable venue for publication, quite the opposite. It does imply that a large amount of academic discussion within the research area is held on a single venue, making the research area less versatile and inclusive. The summary of the results is presented in Table 12.

TABLE 12 Theoretical implications

Theoretical relations of Primary Empirical Conclusions		
Identifier	**Primary Empirical Conclusion**	**Relation to existing research**
PEC1	Explainable AI is significant research focus on the study field of AI Ethics. From the empiric research papers published after 2012, 36.2% is related to Explainable AI.	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 (Jobin et al. 2019)
PEC2	In the study field of Explainable Ethical AI, the most common type of empiric research is to study a novel technique that can solve a computational challenge.	Novel. The lack of validation study shows the freshness of the research field.
PEC3	The human perspective towards Explainable AI is not well-known. There is a lack of research about the practitioners' and user's expectations and attitudes towards Explainable AI.	Corresponding. The same challenge was noted in previous research of Adadi and Berrada (2018)
PEC4	Industrial implementation of Explainable AI is not profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of Explainable AI.	Corresponding. Related to PEC3 and to observations of Vakkuri et al, (2020) in the research area of AI ethics in general.
PEC5	Explainable AI researchers are interested in real-world applications, not only technical aspects of the topic. The empiric	Contradicting. Even though black box research is distancing from the real-world problems (Rudin, 2019) XAI research is

	research area of Explainable AI has a close connection to real-world problems when the research is related to AI ethics.	close to real-world problems.
PEC6	In the corpus of explainable ethical AI, the publication venue of empirical research is monotonic, with 43,4% of the papers published in one conference.	Contradicting. AI and AI ethics are cross-disciplinary research areas studied in several domains (Russell & Norvig, 1994; Vakkuri & Abrahamsson, 2018). Still, the publication of empiric research related to XAI is often published in the same venue.

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

6.2 Practical Implication

Some of the PEC's only had a clear theoretical contribution; hence they are not analyzed by their relevance to practitioners. The research field has a close connection to the real-world by contributing to social issues and using real-world data (PEC5). The research provides knowledge and perspective to regulators and communicators by contributing and tiding the research into 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).

On the other hand, as the research field is new and emerging, the lack of practical implementation is visible (PEC3 and PEC4). The lack of research on attitudes, expectations, or needs of users and practitioners might distant the solutions from the practical needs. 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 is unknown, as well as the expectations or interest of business decision-makers. As long as the decision-makers do not understand XAI's need, the practical implementation in businesses is not likely to happen on a bigger scale (PEC4). The summary of results is presented in Table 13.

TABLE 13 Practical Implications

Practical Implications of Primary Empirical Conclusions		
Identifier	**Primary Empirical Conclusion**	**Practical implications**
PEC2	In the study field of Explainable Ethical AI, the most common type of empiric research is to study a novel technique	The field is still in the research phase, and the practical implementation and validation are missing. There are several open-source

	that can solve a computational challenge.	solution proposals modifiable to fit the company needs. There is significant potential for collaboration between academia and industry.
PEC3	The human perspective towards Explainable AI is not well-known. There is a lack of research about the practitioners' and user's expectations and attitudes towards Explainable AI.	The solutions proposed in the research papers might not have practical implementation potential due to the lack of understanding the practical needs. The regulation of the field is challenging without the understanding of expectations and needs.
PEC4	Industrial implementation of Explainable AI is not profoundly studied in the research field of AI ethics. There is a research gap in the managerial perspective and the business implications of Explainable AI.	It is required to understand if and how the XAI solutions are needed and understood by business decision-makers to find the best solutions and enable practical implementation.
PEC5	Explainable AI researchers are interested in real-world applications, not only technical aspects of the topic. The empiric research area of Explainable AI has a close connection to real-world problems when the research is related to AI ethics.	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.

In conclusion, the analyzes of practical implementation revealed the potential of even closer collaboration between practitioners and academia. On the other hand, the research gap of understanding the practitioners, users, or business decision-makers can harm the practical implementation of XAI solutions. Overall, more research is required to move forward.

7 CONCLUSIONS

In this paper, the Systemic Mapping Study method was utilized to visualize how Explainable AI is researched in the field of AI ethics. SMS was chosen to have a broader perspective with AI ethics to have a more profound understanding of the research area and the role of XAI in the research area. The expected findings included mapping of the covered topic and analysis on when, how, and why the research has done to reveal potential research gaps. This chapter concludes the findings of this study, discusses the limitations, and proposes future research areas.

7.1 Answer to Research Question

The research question was *What is the role of XAI in the research field of AI ethics?* To answer the question, it was required to understand what is researched in AI ethics and how XAI is emphasized in the research field. Besides, it was required to understand what, how, when, and why XAI was researched. Because this paper focused on practical solutions and implementation, the research was narrowed to empirical papers. To answer these questions, the research question was divided into sub-questions:

[R1] What is researched in the AI ethics research field with empiric evidence?
[R2] What is the current state of XAI in the research field of AI ethics?
[R3] What are the research gaps in the field?

Next, each of the questions is answered based on the data and analyses of SMS.

7.1.1 What is researched in the AI ethics research field with empiric evidence?

This paper is interested in XAI's practical implications; hence, the research was narrowed to empiric papers. The short analyzes of the dataset of empiric research in AI ethics (Chapter 3.4) revealed that only 20% of the AI Ethics papers use empiric material. Overall, the AI ethics research is rather theoretic.

Empiric research grew significantly in 2018. In the same year, the topic became popular in media and public discussion. In this paper, the reason for the growth is not studied. Since 2018 the empiric research has kept on growing every year. Similarly, the research focus in XAI grew significantly in 2018 and has kept on growing since.

There were specific emerging themes in the sample of empiric research (n=213), such as the emphasis on autonomous vehicles, Human-Robot Interaction, and health and care robotics. 36.19% of the sample was contributing to issues related to XAI. The percentage includes papers with a partial or a marginal contribution to XAI. Overall, the interest in XAI and related issues is a significant area in AI ethics research, especially in empiric papers.

7.1.2 What is the current state of XAI in the research field of AI ethics?

In chapter 5, the research area was mapped to visualize and analyze the papers' focus, research type, contribution, and pertinence. Also, the annual changes and publication venues were analyzed. Based on the analysis, the Primary Empirical Conclusions were noted and mirrored to existing research in chapter 6. Also, practical contributions were proposed.

In conclusion, XAI research's current state is close to real-world problems, published in the last three years, and often proposing a novel computational solution to technical problems. The research is published in both conferences and journals. The venue of publication is monotonous, which might harm the diversity and inclusivity of the research area. The research area is growing, and each year there are more papers with full focus on XAI.

XAI is still mainly interpreted as an academic challenge, even though transparency issues are often emphasized in companies' or institutions' ethical principles for AI development. The majority of the papers were interested more in the technical or design perspective of the problem than in the practical challenges in implementation. Only one paper studied the current state of industrial implementation of AI ethics, and none of the papers studied the industrial implementation of XAI.

7.1.3 What are the research gaps in the field?

Even though the research area is close to real-world problems, it is still theoretical and lacks the implementation in practice or the research of the implementation. There was a lack of understanding of the users' and practitioners' expectations, needs, and attitudes towards XAI. There was no research on the managerial perspective of XAI. A more profound understanding of the current implementation level is needed to ensure that the research has value for practitioners.

The SMS also revealed a research gap with a profound understanding of AI ethics's research area. This paper only touched the surface of the AI ethics field with empiric material, and the theoretical research of AI ethics has not been studied. A better understanding of the research area could reveal research gaps, and it could visualize what topics are well covered in the overall corpus, what type of research is done, and for what purposes. The practical implementation of AI ethics also requires more studies from different industries and public sectors.

7.2 Limitations

A common bias that systematic reviews suffer is that positive outcomes are more likely to be published than negative ones (Kitchenham and Charters, 2007). Especially in the corpus of empiric research, this might 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 (Kitchenham and Charters, 2007) used in this paper.

The research question's framing had two limitations to the study—first, the challenge with complexity with terms. As the focus of this paper is to understand the research field of AI ethics and the role of XAI in the field, this paper provides the mapping to this specific viewpoint. This viewpoint has its challenges, as this definition leaves out all the research papers with a focus on AI's interpretability without clear visible relation to ethical concerns. To understand the full corpus related to XAI, a separate mapping is needed. Even though XAI is closely related to ethical issues, it is also studied independently. A second limitation with research questions was that the final form of questions was defined during the SMS, which challenged the literature and inclusion process, making it a bit more labor-intensive than expected. The overall goal and main focus stayed the same; hence, the definition did not compromise the research's accuracy.

A challenge with the scope was the unexpectedly large sample size from the primary search, which made the literature search and inclusion process labor-intensive and reduced accuracy. The time invested in each paper had to be cut to a minimum. To ensure the quality and manageable workload, the literature search and inclusion process were divided with two student researchers and were closely supervised by the JYU AI Ethics Lab.

During the primary search, some limitations were faced. Each database was screened, starting from the oldest papers to track its potential changes during the screening process. In Scopus, only the 2000 first articles were allowed to be screened without changing the settings, which required to screen the articles with year range starting with the 2000 oldest and then continue towards the newest. There is a possibility that some of the articles were missed in the screening due to this re-organizing and inaccurate date information. In ProQuest, similar challenges were faced, which required re-ordering the search results. Also, 200 hits in the ProQuest database were excluded from the literature search due to technical problems in ProQuest.

After the primary search, the sample size was larger than expected, which limited the amount of attention dedicated to each paper during the screening process. In other SMSes, the initial take-in from separate databases has been significantly lower; 1062 papers (Vakkuri & Abrahamson, 2018), 1769 papers

(Paternoster et al., 2014) or 2081 papers (Belmonte, Morales, & Fernández-Caballero, 2019). The largest sample in the SMS papers benchmarked for this paper was 5082 papers, which was computationally analyzed (Petersen, Vakkalanka, and Kuzniarz, 2015).

Due to the large sample, the literature search and inclusion processes were conducted mostly by one viewer per paper; hence there is a chance of humane mistakes and false classification during the screening process. If the screener felt uncertain with the paper, the paper was tagged, and another screener provided a second opinion. That ensures a better quality of the paper. Out of the full sample of 1935 papers, two viewers evaluated 150 papers during the second screening round. During the final screening round, 22 out of 213 papers were evaluated by two viewers. The included papers after each screening round were re-evaluated during the screening rounds. The papers excluded or misclassified as idea papers during the first and second screening round were not further evaluated, increasing the possibility of missing a suitable paper from the final study due to manual labeling failure.

There were some limitations in the classification process. The classification can be challenging for undergraduates who are not confident working with titles and abstracts of the papers (Budgen et al., 2008). The main limitation with the keywording and classification was the author's inexperience as a researcher. The classification process was highly opinion based which impairs the quality and liability of the study. To ensure the research quality, the classification schema was presented and evaluated by two viewers, but the classification was performed alone. If the classified material is used in future studies, perhaps re-evaluating the sample and classification is needed before utilizing it.

7.3 Future Research

There is potential to continue the SMS composed in this paper to gain a more in-depth understanding of the AI ethics research field. There is no profound mapping of the research field, and this SMS could provide a base for future research. The literature search and inclusion process were performed with clear guidelines, disciplinary following a stringent search process, which enables the future use for the research material (Kitchenham et al., 2011). Future research requires updating the material, and to increase the quality. Snowballing of the primary study references could reveal more fitting papers.

The SMS revealed research gaps in the existing corpus. There is a need to study how humans perceive XAI, and what are they expecting from XAI systems, or do they even value them. That knowledge could guide the research area to look for solutions that are needed. Perhaps cross-disciplinary research between computer scientists and humanists could provide exciting insights into the field.

Another research gap was the lack of industrial implementation. There were no studies of the current state of implementation outside software engineering, and only one study focusing on the implementation of AI ethics in companies in software engineering. The research field could benefit the knowledge of the current state in practical implementation if there are any. Furthermore, how and who in the companies is now managing the issues with XAI.

Future research is needed to understand the managerial perspective of transparent systems in companies using AI solutions. The top managers are the final decision-makers and accountable for their products' actions, and they are the gatekeepers for funding for development. To ensure the solutions proposed in papers to be implemented in practice, it is required to understand what business decision-makers want and where they are ready to invest.

REFERENCES

AAAI/ACM. (2017). AAAI/ACM Conference on AI, Ethics, and Society (AIES). Retrived 27.8.2020 from https://www.aies-conference.com/2018/

ACM FAccT Conference. (27.8.2020). FAT* Conference. Retrieved 27.8.2020 from https://facctconference.org/2018/index.html

Abeywickrama, D. B., Cirstea C. and Ramchurn, S. D. (2019). Model Checking Human-Agent Collectives for Responsible AI, *28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1-8.* New Delhi, India.

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

Addis, C. and Kutar, M. (2019). AI Management An Exploratory Survey of the Influence of GDPR and FAT Principles, in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, pp. 342-347*. Leicester, United Kingdom.

Ajmeri, N., Guo, H., Murukannaiah, P. K., and Singh, M. P. (2020). Elessar: Ethics in Norm-Aware Agents. *In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20). 16–24.* Richland, SC.

Ala-Pietilä, P., Bauer, W., Bergmann, U., Beliková, M., Bonefeld-Dahl, C., Bonnet, Y., … Yeung, K. (2019). *ETHICS GUIDELINES FOR TRUSTWORTHY AI*. Bryssels: AI HLEG, European Commission.

Ali, J., Zafar, M.B., Singla, A., and Gummadi, K.P. (2019). Loss-Aversively Fair Classification. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 211–218.* New York, USA.

Alpaydin, E. (2014). Introduction to Machine Learning, Third edition, *MIT Press*. Retrived from http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=3339851.

Alpaydin, E. (2016). Machine Learning : The New AI, *MIT Press*. Retrived from http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=4714219.

Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., and Rus, D. (2019). Uncovering and Mitigating Algorithmic Bias through Learned Latent

Structure. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 289–295.* New York, USA.

Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A., and Aïvodji, U. (2019). Fairwashing: the risk of rationalization. *International Conference on Machine Learning.*

Asimov, I. (1942). Runaround. *Street & Smith.*

Awad, E., Dsouza, S., Kim, R. *et al*. (2018). The Moral Machine experiment. *Nature, 563***,** 59–64.

Babu, M., and Pushpa, S. (2018). An Efficient Discrimination Prevention and Rule Protection Algorithms Avoid Direct and Indirect Data Discrimination in Web Mining. *International Journal of Intelligent Engineering and Systems.*

Balachander, T., Batra, A. K., and Choudhary, M. (2020). Machine Learning Pipeline for an Improved Medical Decision Support. *International Journal of Advanced Science and Technology. Vol. 29, No. 6,* 2632-2640.

Barn, B. S. (2019). Mapping the public debate on ethical concerns: Algorithms in mainstream media. *Journal of Information, Communication & Ethics in Society, 17*(1), 38-53.

Belmonte, L. M., Morales, R., and Fernández-Caballero, A. (2019). Computer vision in autonomous unmanned aerial Vehicles—A systematic mapping study. *Applied Sciences, 9*(15).

Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. (2019). Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 453–459.* New York, USA.

Brandão, M., Jirotka, M., Webb, H, and Luff, P. (2020). Fair navigation planning: A resource for characterizing and designing fairness in mobile robots, *Artificial Intelligence, Volume 282, 103259.*

Bremner, P., Dennis, L. A., Fisher, M., and Winfield, A. F. (2019). On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots. *In Proceedings of the IEEE, vol. 107, no. 3*, 541-561.

Brundage, M. (2014). Limitations and risks of machine ethics, *Journal of Experimental & Theoretical Artificial Intelligence, 26:3*, 355-372.

Brunk, J., Mattern, J., and Riehle, D. M. (2019). Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. *IEEE 21st Conference on Business Informatics (CBI).* 429-435.

Bryson, J. J. (2019). The Past Decade and Future of AI's Impact on Society. In Towards a New Enlightenment? *A Transcendent Decade, (Vol. 11).*

Budgen, D., Turner, M., Brereton, P., & Kitchenham, B.A. (2008). Using Mapping Studies in Software Engineering. *PPIG.*

Bynum, T.W. (2001). Computer ethics: Its birth and its future. *Ethics and Information Technology 3*, 109–112

Caliskan, A, Bryson, JJ & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, vol. 356, no. 6334,* 183-186.

Calmon, F. d. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N. and Varshney, K. R. (2018). Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis. *IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 5,* 1106-1119.

Cardoso, R. L., Meira Jr., W., Almeida, V., and Zaki, M. J. (2019). A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 437–444.* New York, USA.

Celis, D., and Rao, M. (2019). Learning Facial Recognition Biases through VAE Latent Representations. *In Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia (FAT/MM '19). 26–32.* New York, USA.

Chen, V. X., and Hooker, J. N. (2020). A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 221–227.* New York, USA.

Clavell, G. G., Zamorano, M. M., Castillo, C., Smith, O., and Matic, A. (2020). Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 265–271.* New York, USA.

Cortés E. C., and Ghosh, D. (2020). An Invitation to System-wide Algorithmic Fairness. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 235–241.* New York, USA.

Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. (2019). Fair Transfer Learning with Missing Protected Attributes. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 91–98.* New York, USA.

Crockett, K., Goltz, S., Garratt, M., and Latham, A. (2019). Trust in Computational Intelligence Systems: A Case Study in Public Perceptions. *2019 IEEE Congress on Evolutionary Computation (CEC),* pp. 3227-3234.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved 12.8.2020 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed 14 May 2019.

Dignum, V. (2017). Responsible Artificial Intelligence: Designing AI for Human Values. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).* Melbourne, Australia.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). 67–73.* New York, USA.

Ehsan, U., Harrison, B., Chan, L., and Riedl, M. O. (2018). Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). 81–87.* New York, USA.

Evgeniou, T., Hardoon D. R., and Ovchinnikov, A. (2020). What Happens When AI is Used to Set Grades? *Harward Business Review.* Retrieved from https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades

Faraj, S. & Sambamurthy, V. (2006). Leadership of Information Systems Development Projects. *IEEE Transactions on Engineering Management, 53(2),* 238–249.

Ferreira J.J., and Monteiro M.S. (2020). What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. *HCII 2020. Lecture Notes in Computer Science, vol 12201. Springer, Cham.*

Fieser, J, (2020). Ethics. *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002, University of Tennessee at Martin.

Flexer, A., Dörfler, M., Schlüter, J., and Grill, T. (2018). Hubness as a Case of Technical Algorithmic Bias in Music Recommendation. *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1062-1069, doi: 10.1109/ICDMW.2018.00154.

Floridi, L. (2009). Foundations of Information Ethics. *The Handbook of Information and Computer Ethics (1-23).*

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. (2019). Counterfactual Fairness in Text Classification through Robustness. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 219–226.* New York, USA.

Garousi, V., Felderer, M., § Mäntylä, M.V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering, *Information and Software Technology, Volume 106,* 101-121.

Goel, N., and Faltings, B. (2019). Crowdsourcing with Fairness, Diversity and Budget Constraints. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 297–304*. New York, USA.

Green, B., and Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact. 3, CSCW*, Article 50, 24 pages.

Grgic-Hlaca, N., Zafar, M., Gummadi, K., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *AAAI.*

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *In Proceedings of the 2018 World Wide Web Conference (WWW '18). 903–912.* Republic and Canton of Geneva, CHE.

He, Y., Burghardt, K., and Lerman, K. (2020). A Geometric Solution to Fair Representations. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 279–285.* New Yor, USA.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). 181–190.* New York, USA.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. (2018). Ethical Challenges in Data-Driven Dialogue Systems. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).* New York, USA.

Hind, M., Wei, D., Campbell, M., Codella, N. C. F., Dhurandhar, A., Mojsilović, A., Ramamurthy, K. N., and Varshney, K. R. (2019). TED: Teaching AI to Explain its Decisions. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 123–129.* New York, USA.

Holm, E. A. (2019). In defense of the black box. *Science, 364(6435),* 26-27.

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, First Edition.* (IEEE, 2019). Retrieved from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html

Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., and Sycara, K. (2018). Transparency and Explanation in Deep Reinforcement Learning Neural Networks. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). 144–150.* New York, USA.

Jo, E.S., and Gebru, T. (2020). Lessons from archives: strategies for collecting sociocultural data in machine learning. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). 306–316.* New York, USA.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1–11,

Karpati, D., Najjar, A., and Ambrossio, D. A. (2020). Ethics of Food Recommender Applications. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 313–319.* New York, USA.

Kitchenham, B. & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Technical Report EBSE 2007-001.

Kitchenham, B., Budgen, D., and Brereton, P. (2011). Using mapping studies as the basis for further research - A participant-observer case study. *Information & Software Technology. 53. 638-651.*

Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 247–254.* New York, USA.

Lai, V., and Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). 29–38.* New York, USA.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). Faithful and Customizable Explanations of Black Box Models. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 131–138.* New York, USA.

Lakkaraju, H., and Bastani, O. (2020). ”How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 79–85*. New York, USA.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553),* 436-444.

Lux, T. C. H., Nagy, S., Almanaa, M., Yao, S., and Bixler, R. (2019). A Case Study on a Sustainable Framework for Ethically Aware Predictive Modeling. *2019 IEEE International Symposium on Technology and Society (ISTAS),* 1-7.

Madaio, M. A., Stark, L., Vaughan, J. W., and Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). 1–14.* New York, USA.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Ithaca: Cornell University Library, arXiv.org.*

Ministry of economic affairs and employment of Finland (2020). *Ethical Challenge.* Retrevied 31.6.2020 from https://www.tekoalyaika.fi/en/background/ethics/

Mitchell, T.M. (1980). *The Need for Biases in Learning Generalizations.* Rutgers CS tech report CBM-TR-117.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). 220–229.* New York, USA.

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T., and Morgenstern, J. (2020). Diversity and Inclusion Metrics in Subset Selection. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 117–123.* New York, USA.

Moor, J. (2011). The Nature, Importance, and Difficulty of Machine Ethics. In M. Anderson & S. Anderson (Eds.), *Machine Ethics*, 13-20.

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci Eng Ethics 26*, 2141–2168.

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. (2019). Active Fairness in Algorithmic Decision Making. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 77–83.* New York, USA.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *New York: Crown Publishers.* 272p.

Orr, W., and Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners, *Information, Communication & Society, 23:5*, 719-735.

Ouchchy, L., Coin, A. & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & Soc 35,* 927–936.

Paternoster, Nicolò & Giardino, Carmine & Unterkalmsteiner, Michael & Gorschek, Tony & Abrahamsson, Pekka. (2014). Software Development in Startup Companies: A Systematic Mapping Study. *Information and Software Technology*. 56.

Petersen, K., Feldt, R., Mujtaba, S. and Mattsson, M. (2008). Systematic mapping studies in software engineering. *In Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering (EASE'08).* BCS Learning & Development Ltd., Swindon, GBR, 68–77.

Petersen, K., Vakkalanka, S., Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: *An update, Information and Software Technology*, *Volume 64,* 1-18.

Paternoster, N., Giardino C., Unterkalmsteiner, M., Gorschek, T., Abrahamsson, P. (2014). Software development in startup companies: A systematic mapping study, *Information and Software Technology, Volume 56, Issue 10,* 1200-1218.

Radovanović, S., Petrović, A., Delibašić, B., & Suknović, M. (2019). Making hospital readmission classifier fair – what is the cost?. *Proceedings of the Central European Conference on Information and Intelligent Systems.* Varaždin, Croatia.

Raff, E., Sylvester, J., and Mills, S. (2018). Fair Forests: Regularized Tree Induction to Minimize Model Bias. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). 243–250.* New York, USA.

Raji, I. D., and Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 429–435.* New York, USA.

Rossi, F., Ala-Pietilä, P., Bauer, W., Bergmann, U., Beliková, M., Bonefeld-Dahl, C., … Yeung, K. (2019). *A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES.* Bryssels: European Comission.

Rubel, A., Castro, C. & Pham, A. (2019). Agency Laundering and Information Technologies. *Ethic Theory Moral Prac* 22, 1017–104.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell 1,* 206–215.

Russell, S. & Norvig, P. (1994). Artificial Intelligence: A Modern Approach. *Prentice Hall.*

Russell, S.J., and Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3nd. ed.). *Pearson Education.*

Samek, W., Wiegand, T., & Müller, K. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services. 1.* 1-10.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y. (2019). How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 99–106.* New York, USA.

Schneider, J., & Handali, J. (2019). Personalized explanation for machine learning: A conceptualization. *27th European Conference on Information Systems (ECIS 2019).* Stockholm-Uppsala, Sweden.

Schelter, S., He, Y., Khilnani, J., Stoyanovich, J. (2020). FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT).* Copenhagen, Dennmark.

Shaw, M. (2003). Writing good software engineering research papers, *25th International Conference on Software Engineering, IEEE Computer Society. 726-736.* Buenos Aires, Argentina.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., and O'Brien, C. (2020). "The human body is a black box": supporting clinical decision-making with deep learning. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). 99–109.* New York, USA.

Shank, D. B., and DeSanti, A., (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior. Volume 86*, 401-411.

Sharma, S., Henderson, J., and Ghosh, J. (2020a). CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 166–172.* New York, USA.

Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D. Muthusamy, V., and Varshney, K. R. (2020b). Data Augmentation for Discrimination Prevention and Bias Disambiguation. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 358–364.* New York, USA.

Shulman E., and Wolf, L. (2020). Meta Decision Trees for Explainable Recommendation Systems. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). A365–371.* New York, USA.

Sivill, T. (2019). Ethical and Statistical Considerations in Models of Moral Judgments. *Frontiers in Robotics and AI. Volume 6*, 39.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). 180–186.* New York, USA.

Srinivasan, R. & Chander, A. (2019). Understanding Bias in Datasets using Topological Data Analysis. In *AISafety@IJCA*. USA.

Srivastava, B., and Rossi, F. (2018). Towards Composable Bias Rating of AI Services. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).* 284–289. New York, USA.

Templier, M and Paré, G. (2015). A Framework for Guiding and Evaluating Literature Reviews, *Communications of the Association for Information Systems: Vol. 37, Article 6.*

Teso, S., and Kersting, K. (2019). Explanatory Interactive Machine Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). 239–245.* New York, USA.

Ustun, B., Liu, Y., & Parkes, D. (2019). Fairness without Harm: Decoupled Classifiers with Preference Guarantees. *Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:6373-6382.* Long Beach, USA.

van den Hoven, J. (2009). Moral Methodology and Information Technology. *The Handbook of Information and Computer Ethics* (eds K.E. Himma and H.T. Tavani).

Vakkuri, V. & Abrahamsson, P. (2018). The Key Concepts of Ethics of Artificial Intelligence, *IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 1-6,* Stuttgart, Gemany.

Vakkuri, V., Kemell, K. and Abrahamsson, P. (2019). Ethically Aligned Design: An Empirical Evaluation of the RESOLVEDD-Strategy in Software and Systems Development Context, *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 46-50.* Kallithea-Chalkidiki, Greece.

Vakkuri, V., Kemell, K., Kultanen, J. and Abrahamsson, P. (2020). The Current State of Industrial Practice in Artificial Intelligence Ethics. *In IEEE Software, vol. 37, no. 4,* 50-57.

Vanderelst, D., Willems, J. (2019). Can We Agree on What Robots Should be Allowed to Do? An Exercise in Rule Selection for Ethical Care Robots. *Int J of Soc Robotics*.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*.

Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). 440, 1–14.* New York, USA.

Vetrò, A., Santangelo, A., Beretta, E. and De Martin, J.C. (2019). AI: from rational agents to socially responsible agents, *Digital Policy, Regulation and Governance*, *Vol. 21 No. 3,* 291-304.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, Volume 31, Number 2 Spring 2018,* 842-861

Wieringa, R., Maiden, N., Mead, N. and Rolland, C. (2006). Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. *Requirements Eng 11***,** 102–107.

Webb, H., Patel, M., Rovatsos, M., Davoust, A., Ceppi, S., Koene, A., . . . Cano, M. (2019). It would be pretty immoral to choose a random algorithm. *Journal of Information, Communication & Ethics in Society, 17(2),* 210-228.

Wolf, L., Galanti, T., and Hazan, T. (2019). A Formal Approach to Explainability. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19).* New York, USA.

Wouters, N., Kelly, R., Velloso, E., Wolf, K., Ferdous, H. S., Newn, J., Joukhadar, Z., and Vetere, F. (2019). Biometric Mirror: Exploring Ethical Opinions towards Facial Analysis and Automated Decision-Making. *In Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19). 447–461.* New York, USA.

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., and Miklau, G. (2018). A Nutritional Label for Rankings. *In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). 1773–1776.* New York, USA.

Yilmaz, L., and Sivaraj, S. (2019). A Cognitive Architecture for Verifiable System Ethics via Explainable Autonomy, *2019 IEEE International Systems Conference (SysCon), 1-8,* Orlando, FL, USA.

Zhang, B.H., Lemoine, B., and Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). 335–340.* New York, USA.

Zhou, J., and Chen, F. (2018). DecisionMind: revealing human cognition states in data analytics-driven decision making with a multimodal interface. *Journal on Multimodal User Interfaces 12*, 67–76.

APPENDIX I – DATASET OF EMPIRIC PAPERS N=212

This dataset has not been inspected for grey or black literature and it might include short papers, pre-publications or other papers that did not meet the academic quality requirements implemented to the final dataset presented in Table 8.

Authors	Title	Year	DOI
Abeywickrama, D. B.; Cirstea, C.; Ramchurn, S. D.	Model Checking Human-Agent Collectives for Responsible AI	2019	10.1109/RO-MAN46459.2019.8956429
Addis, C.; Kutar, M.	AI Management An Exploratory Survey of the Influence of GDPR and FAT Principles	2019	10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00102
Addison, A.; Bartneck, C.; Yogeeswaran, K.	Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots	2019	10.1145/3306618.3314272
Aïvodji U., Arai H., Fortineau O., Gambs S., Hara S., Tapp A.	Fairwashing: The risk of rationalization	2019	
Ajmeri, N.; Guo, H.; Murukannaiah, P.K.; Singh, M.P.	Elessar: Ethics in Norm-Aware Agents	2020	10.5555/3398761.3398769
Al Barghuthi, N. B.; Said, H.	Readiness, Safety, and Privacy on Adopting Autonomous Vehicle Technology: UAE Case Study	2019	10.1109/ITT48889.2019.9075090
Al Nahian, S.; Frazier, S.; Riedl, M.; Harrison, B.	Learning Norms from Stories: A Prior for Value Aligned Agents	2020	10.1145/3375627.3375825
Ali, J.; Zafar, M.B.; Singla, A.; Gummadi, K.P.	Loss-Aversively Fair Classification	2019	10.1145/3306618.3314266
Alkoby, S.; Rath, A.; Stone, P.	Teaching Social Behavior through Human Reinforcement for Ad hoc Teamwork - The STAR Framework: Extended Abstract	2019	
Amigoni, F.; Schiaffonati, V.	Ethics for Robots as Experimental Technologies: Pairing Anticipation with Exploration to Evaluate the Social Impact of Robotics	2018	10.1109/MRA.2017.2781543
Amini, A.; Soleimany, A.P.; Schwarting, W.; Bhatia, S.N.; Rus, D.	Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure	2019	10.1145/3306618.3314243

Anderson, M.; Anderson, S. L.; Berenz, V.	A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm	2019	10.1109/JPROC.2018.2840045
Anderson, M.; Anderson, S.L.	Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm	2015	10.1108/IR-12-2014-0434
Aroyo, A. M.; Kyohei, T.; Koyama, T.; Takahashi, H.; Rea, F.; Sciutti, A.; Yoshikawa, Y.; Ishiguro, H.; Sandini, G.	Will People Morally Crack Under the Authority of a Famous Wicked Robot?	2018	10.1109/ROMAN.2018.8525744
Avin, S.; Gruetzemacher, R.; Fox.	Exploring AI Futures Through Role Play	2020	10.1145/3375627.3375817
Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I.	The Moral Machine experiment	2018	10.1038/s41586-018-0637-6
Awad E., Levine S., Kleiman-Weiner M., Dsouza S., Tenenbaum J.B., Shariff A., Bonnefon J.-F., Rahwan I.	Drivers are blamed more than their automated cars when both make mistakes	2020	10.1038/s41562-019-0762-8
Babu M.C., Pushpa S.	An efficient discrimination prevention and rule protection algorithms avoid direct and indirect data discrimination in web mining	2018	10.22266/ijies2018.0831.21
Balachander T., Batra A.K., Choudhary M.	Machine learning pipeline for an improved medical decision support	2020	
Banks J.	A perceived moral agency scale: Development and validation of a metric for humans and social machines	2019	10.1016/j.chb.2018.08.028
Barn, B. S.	Mapping the public debate on ethical concerns: algorithms in mainstream media	2019	10.1108/JICES-04-2019-0039.
Battistuzzi, L.; Sgorbissa, A.; Papadopoulos, C.; Papadopoulos, I.; Koulouglioti, C.	Embedding Ethics in the Design of Culturally Competent Socially Assistive Robots	2018	10.1109/IROS.2018.8594361
Bergmann L.T., Schlicht L., Meixner C., König P., Pipa G., Boshammer S., Stephan A.	Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making	2018	10.3389/fnbeh.2018.00031

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Woodruff, A.; Luu, C.; Kreitmann, P.; Bischof, J.; Chi, E.H.	Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements	2019	10.1145/3306618.3314234
Bigman Y.E., Gray K.	People are averse to machines making moral decisions	2018	10.1016/j.cognition.2018.08.003
Bonnefon, J-F; Shariff, A; Rahwan, I	The social dilemma of autonomous vehicles	2016	10.1126/science.aaf2654
Brandão M., Jirotka M., Webb H., Luff P.	Fair navigation planning: A resource for characterizing and designing fairness in mobile robots	2020	10.1016/j.artint.2020.103259
Bremner, P.; Dennis, L. A.; Fisher, M.; Winfield, A. F.	On Proactive, Transparent, and Verifiable Ethical Reasoning for Robot	2019	10.1109/JPROC.2019.2898267
Brunk, J.; Mattern, J.; Riehle, D. M.	Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems	2019	10.1109/CBI.2019.00056
Burton, E.;Goldsmith, J.;Koenig, S.;Kuipers, B.;Mattei, N.;Walsh, T.	Ethical Considerations in Artificial Intelligence Courses	2017	10.1609/aimag.v38i2.2731
Bussmann B.; Heinerman J.; Lehman J.	Towards empathic deep q-learning	2019	
Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind	Semantics derived automatically from language corpora contain human-like biases	2017	10.1126/science.aal4230
Calmon, F. d. P.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; Varshney, K. R.	Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis	2018	10.1109/JSTSP.2018.2865887
Cave, S.; Coughlan, K.; Dihal, K.	“Scary Robots”: Examining Public Responses to AI	2019	10.1145/3306618.3314232
Celis, D.; Rao, M.;	Learning Facial Recognition Biases through VAE Latent Representations	2019	10.1145/3347447.3356752
Censi, A.; Slutsky, K.; Wongpiromsarn, T.; Yershov, D.; Pendleton, S.; Fu, J.; Frazzoli, E.	Liability, Ethics, and Culture-Aware Behavior Specification using Rulebooks	2019	10.1109/ICRA.2019.8794364
Chan, L.; Doyle, K.; McElfresh, D.; Conitzer, V.; Dickerson, J.P.; Borg, J.S.; Sinnott-Armstrong, W.	Artificial Artificial Intelligence: Measuring Influence of AI “Assessments” on Moral Decision-Making	2020	10.1145/3375627.3375870

Chatterjee S.	Impact of AI regulation on intention to use robots: From citizens and government perspective	2019	10.1108/IJIUS-09-2019-0051
Chen, V.; Hooker, J.N.	A Just Approach Balancing Rawlsian Leximax Fairness and Utilitarianism	2020	10.1145/3375627.3375844
Cheon, E; Makoto Su, N.	Integrating Roboticist Values into a Value Sensitive Design Framework for Humanoid Robots	2016	10.1109/HRI.2016.7451775
Chernyak, N.; Gary, H. E.	Children's Cognitive and Behavioral Reactions to an Autonomous Versus Controlled Social Robot Dog	2016	10.1080/10409289.2016.1158611
Coston, A.; Ramamurthy, K.N.; Wei, D.; Varshney, K.R; Speakman, S.; Mustahsan, Z.; Chakraborty, S.	Fair Transfer Learning with Missing Protected Attributes	2019	10.1145/3306618.3314236
Crockett, K.; Goltz, S.; Garratt, M.; Latham, A.	Trust in Computational Intelligence Systems: A Case Study in Public Perceptions	2019	10.1109/CEC.2019.8790147
Cruz Cortés, E.; Ghosh, D.	An Invitation to System-wide Algorithmic Fairness	2020	10.1145/3375627.3375860
D'Aquin M., Troullinou P., O'Connor N.E., Cullen A., Faller G., Holden L.	Towards an "Ethics by Design" Methodology for AI Research Projects	2018	10.1145/3278721.3278765
Dang,L. M.; Min,K.; Lee,S.; Han,D.; Moon,H.	Tampered and Computer-Generated Face Images Identification Based on Deep Learning	2020	10.3390/app10020505
Daniel Karpati, Amro Najjar, and Diego Agustin Ambrossio	Ethics of Food Recommender Applications	2020	10.1145/3375627.3375874
Danielson, P.	Surprising judgments about robot drivers: Experiments on raising expectations and blaming humans	2015	10.5324/eip.v9i1.1727
De Greef, T., Mohabir, A., Van Der Poel, I., Neerincx, M.	sCEthics: Embedding ethical values in cognitive engineering	2013	10.1145/2501907.2501935
Dietrich M., Weisswange T.H.	Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios	2019	10.1007/s10676-019-09504-3
Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; Vasserman, L.	Measuring and Mitigating Unintended Bias in Text Classification	2018	10.1145/3278721.3278729

Dolianitis, A.; Chalkiadakis, C.; Mylonas, C.; Tzanis, D.	How Will Autonomous Vehicles Operate in an Unlawful Environment? - The Potential of Autonomous Vehicles for Disregarding the Law	2019	10.1109/MTITS.2019.8883326
Ehsan, U.; Harrison, B.; Chan, L.; Riedl, M.O.	Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations	2018	10.1145/3278721.3278736
Eichenberg C., Khamis M., Hübner L.	The attitudes of therapists and physicians on the use of sex robots in sexual therapy: Online survey and interview study	2019	10.2196/13853
Eicher, B.; Polepeddi, L.; Goel, A.	Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies	2018	10.1145/3278721.3278760
El Khattabi, G.; Haij, O.; Benelallam, I.; Bouyakhf, E.H.	Detection of Unethical Intelligent Agents in Ethical Distributed Constraint Satisfaction Problems	2018	10.1145/3177148.3180083
Ema A., Osawa H., Saijo R., Kubo A., Otani T., Hattori H., Akiya N., Kanzaki N., Kukita M., Komatani K., Ichise R.	Clarifying Privacy, Property, and Power: Case Study on Value Conflict between Communities	2019	10.1109/JPROC.2018.2837045
Fernandes, P.M.; Santos, F.C.; Lopes, M.	Adoption Dynamics and Societal Impact of AI Systems in Complex Networks	2020	10.1145/3375627.3375847
Feroz,I.; Ahmad,N.; Iqbal,M.W.; Main,N.A.; Shahzad, S.K.	People perception of autonomous vehicles: Legal and ethical issues	2019	10.21833/ijaas.2019.05.015
Flexer, A.; Dörfler M.; Schlüter J.; Grill, T.	Hubness as a Case of Technical Algorithmic Bias in Music Recommendation	2018	10.1109/ICDMW.2018.00154
Frank D.-A., Chrysochou P., Mitkidis P., Ariely D.	Human decision-making biases in the moral dilemmas of autonomous vehicles	2019	10.1038/s41598-019-49411-7
Frey W.R., Patton D.U., Gaskell M.B., McGregor K.A.	Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data	2020	10.1177/0894439318788314
Frison, A-K.; Wintersberger, P.; Riener. A.	First Person Trolley Problem: Evaluation of Drivers' Ethical Decisions in a Driving Simulator	2016	10.1145/3004323.3004336
Galdon Clavell, G.; Martín Zamorano, M.; Castillo, C.; Smith, O.; Matic, A.	Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization	2020	10.1145/3375627.3375852

Gao Z., Sun Y., Hu H., Zhang T., Gao F.	Investigation of the instinctive reaction of human drivers in social dilemma based on the use of a driving simulator and a questionnaire survey	2020	10.1080/15389588.2020.1739274
Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E.H.; Beutel, A.	Counterfactual Fairness in Text Classification through Robustness	2019	10.1145/3306618.3317950
Garrett N., Beard N., Fiesler C.	More than "if time allows": The role of ethics in AI education	2020	10.1145/3375627.3375868
Goel, N.; Faltings, B.	Crowdsourcing with Fairness, Diversity and Budget Constraints	2019	10.1145/3306618.3314282
Gogoll J., Uhl M.	Rage against the machine: Automation in the moral domain	2018	10.1016/j.socec.2018.04.003
Goudzwaard, M.; Smakman, M.; Konijn, E. A.	Robots are Good for Profit: A Business Perspective on Robots in Education	2019	10.1109/DEVLRN.2019.8850726
Green, B.; Chen, Y.	The Principles and Limits of Algorithm-in-the-Loop Decision Making	2019	10.1145/3359152.
Grgić-Hlača N., Zafar M.B., Gummadi K.P., Weller A.	Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning	2018	
Grgic-Hlaca, N.; Redmiles, E.M.; Gummadi, K.P.; Weller, A.	Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction	2018	10.1145/3178876.3186138
Guarini, M.	Moral Case Classification and the Nonlocality of Reasons	2013	10.1007/s11245-012-9130-2
Gupta K.P.	Artificial intelligence for governance in india: Prioritizing the challenges using analytic hierarchy process (AHP)	2019	10.35940/ijrte.B3392.078219
Hadfield-Menell, D.; Andrus, M.; Hadfield, G.	Legible Normativity for AI Alignment: The Value of Silly Rules	2019	10.1145/3306618.3314258
He, Y.; Burghardt, K.; Lerman, K.	A Geometric Solution to Fair Representations	2020	10.1145/3375627.3375864
Heidari, H.; Loi, M.; Gummadi, K.P.; Krause, A.	A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity	2019	10.1145/3287560.3287584
Henderson, P.; Sinha, K.; Angelard-Gontier, N.; Ke, N.R.; Fried, G.; Lowe, R.; Pineau, J.	Ethical Challenges in Data-Driven Dialogue Systems	2018	10.1145/3278721.3278777

Hind, M.; Wei, D.; Campbell, M.; Codella, N. C. F.; Dhurandhar, A.; Mojsilović, A.; Ramamurthy, K.N.; Varshney, K.R.	TED: Teaching AI to Explain its Decisions	2019	10.1145/3306618.3314273
Hoorn J.F., Winter S.D.	Here Comes the Bad News: Doctor Robot Taking Over	2018	10.1007/s12369-017-0455-2
Hu Zhang, B.; Lemoine, B.; Mitchell, M.	Mitigating Unwanted Biases with Adversarial Learning	2018	10.1145/3278721.3278779
Insaurralde, C. C.; Blasch, E.	Moral autonomy in decision-making support from avionics analytics ontology	2018	10.1109/ICNSURV.2018.8384902
Iyer, R; Li, Y.; Li, H.; Lewis, M.; Sundar, R.; Sycara, K.	Transparency and Explanation in Deep Reinforcement Learning Neural Networks	2018	10.1145/3278721.3278776
Jackson, R. B.; Williams, T.	Language-Capable Robots may Inadvertently Weaken Human Moral Norms	2019	10.1109/HRI.2019.8673123
Jiang, R.; Chiappa, S.; Lattimore, T.; György, A.; Kohli, P.	Degenerate Feedback Loops in Recommender Systems	2019	10.1145/3306618.3314288
Jo, E.S.; Gebru, T.	Lessons from archives: strategies for collecting sociocultural data in machine learning	2020	10.1145/3351095.3372829
Kahn P. H.; Kanda, T.; Ishiguro, H.; Gill, B. T.; Ruckert, J. H.; Shen S.; Gary, H. E.; Reichert, A. L.; Freier, N. G.; Severson R. L.	Do people hold a humanoid robot morally accountable for the harm it causes?	2012	10.1145/2157689.2157696
Kallioinen N., Pershina M., Zeiser J., Nosrat Nezami F., Pipa G., Stephan A., König P.	Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives	2019	10.3389/fpsyg.2019.02415
Kasenberg, D.; Arnold, T.; Scheutz, M.	Norms, Rewards, and the Intentional Stance: Comparing Machine Learning Approaches to Ethical Training	2018	10.1145/3278721.3278774
Kasenberg, D.; Scheutz, M.	Inverse Norm Conflict Resolution	2018	10.1145/3278721.3278775
Kerr A., Barry M., Kelleher J.D.	Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance	2020	10.1177/2053951720915939

Kim, R.; Kleiman-Weiner, M.; Abeliuk, A.; Awad, E.; Dsouza, S.; Tenenbaum, J.B.; Rahwan, I.	A Computational Model of Commonsense Moral Decision Making	2018	10.1145/3278721.3278770
Korn, O.; Bieber, G.; Fron, C.	Perspectives on Social Robots: From the Historic Background to an Experts' View on Future Developments	2018	10.1145/3197768.3197774
Krafft, P.M.; Young, M.; Katell, M.; Huang, K.; Bugingo, G.	Defining AI in Policy versus Practice	2020	10.1145/3375627.3375835
Lai, V.; Tan, C.	On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection	2019	10.1145/3287560.3287590
Lakkaraju, H.; Bastani, O.	"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations	2020	10.1145/3375627.3375833
Lakkaraju, H.; Kamar, E.; Caruana, R.; Leskovec, J.	Faithful and Customizable Explanations of Black Box Models	2019	10.1145/3306618.3314229
Li, H.; Milani, S.; Krishnamoorthy, V.; Lewis, M.; Sycara, K.	Perceptions of Domestic Robots' Normative Behavior Across Cultures	2019	10.1145/3306618.3314251
Li, J., Zhao, X., Cho, M.-J., Ju, W., Malle, B.F.	From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars	2016	10.4271/2016-01-0164
Li, S.; Zhang, J.; Li, P.; Wang, Y.; Wang, Q.	Influencing Factors of Driving Decision-Making Under the Moral Dilemma	2019	10.1109/ACCESS.2019.2932043
Ljungholm D.P.	The safety and reliability of networked autonomous vehicles: Ethical dilemmas, liability litigation concerns, and regulatory issues	2019	10.22381/CRLSJ11220191
Loreggia A., Mattei N., Rossi F., Venable K.B.	Metric learning for value alignment	2019	
Lux, T. C. H.; Nagy, S.; Almanaa, M.; Yao, S.; Bixler, R.	A Case Study on a Sustainable Framework for Ethically Aware Predictive Modeling	2019	10.1109/ISTAS48451.2019.8937885
Malle B.F.; Scheutz, M.; Arnold, T.; Voiklis, J.; Cusimano, C.	Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents	2015	10.1145/2696454.2696458

Malle, B.F., Scheutz, M., Forlizzi, J., Voiklis, J.	Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot	2016	10.1109/HRI.2016.7451743
Manikonda L., Kambhampati S.	Tweeting AI: Perceptions of lay versus expert twitterati	2018	
Mattei, N.; Saffidine, A.; Walsh, T.	Fairness in Deceased Organ Matching	2018	10.1145/3278721.3278749
Maurer S., Erbach R., Kraiem I., Kuhnert S., Grimm P., Rukzio E.	Designing a guardian angel: Giving an automated vehicle the possibility to override its driver	2018	10.1145/3239060.3239078
Maurice, P.; Allienne, L.; Malaisé, A.; Ivaldi, S.	Ethical and Social Considerations for the Introduction of Human-Centered Technologies at Work	2018	10.1109/ARSO.2018.8625830
McStay A.	Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy	2020	10.1177/2053951720904386
Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach	Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI	2020	10.1145/3313831.3376445
Michael P. Kim, Amirata Ghorbani, and James Zou	Multiaccuracy: Black-Box Post-Processing for Fairness in Classification	2019	10.1145/3306618.3314287
Mitchell, M.; Baker, D.; Moorosi, N.; Denton, E.; Hutchinson, B.; Hanna, A.; Gebru, T.; Morgenstern, J.	Diversity and Inclusion Metrics in Subset Selection	2020	10.1145/3375627.3375832
Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I.D.; Gebru, T.	Model Cards for Model Reporting	2019	10.1145/3287560.3287596
Mordue G., Yeung A., Wu F.	The looming challenges of regulating high level autonomous vehicles	2020	10.1016/j.tra.2019.11.007
Moussawi, S.	User Experiences with Personal Intelligent Agents: A Sensory, Physical, Functional and Cognitive Affordances View	2018	10.1145/3209626.3209709
Nagataki, S.; Ohira, H.; Kashiwabata, T.; Konno, T.; Hashimoto, T.; Miura, T; Shibata M.; Kubota, S.;	Can Morality Be Ascribed to Robot?	2019	10.1145/3335595.3335643
Nomura, T.; Kanda, T.; Yamada, S.	Measurement of Moral Concern for Robots	2019	10.1109/HRI.2019.8673095

Nomura, T.; Otsubo, K.; Kanda, T.	Preliminary Investigation of Moral Expansiveness for Robots	2018	10.1109/ARSO.2018.8625717
Noothigattu R., Gaikwad S.N.S., Awad E., Dsouza S., Rahwan I., Ravikumar P., Procaccia A.D.	A voting-based system for ethical decision making	2018	
Noriega-Campero, A.; Bakker, M.A.; Garcia-Bulle, B.; Pentland, A.	Active Fairness in Algorithmic Decision Making	2019	10.1145/3306618.3314277
O'Leary D.E.	GOOGLE'S Duplex: Pretending to be human	2019	10.1002/isaf.1443
Omari, R., M; Mohammadian, M.	Rule based fuzzy cognitive maps and natural language processing in machine ethics	2016	10.1108/JICES-10-2015-0034
Orr W., Davis J.L.	Attributions of ethical responsibility by Artificial Intelligence practitioners	2020	10.1080/1369118X.2020.1713842
Ouchchy L., Coin A., Dubljević V.	AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media	2020	10.1007/s00146-020-00965-5
Peca, A.; Coeckelbergh, M.; Simut, R.; Costescu, C.; Pintea, S.; David, D.; Vanderborght, B.	Robot Enhanced Therapy for Children with Autism Disorders Measuring Ethical Acceptability	2016	10.1109/MTS.2016.2554701
Peters, D.; Vold, K.; Robinson, D.; Calvo, R. A.	Responsible AI—Two Frameworks for Ethical Design Practice	2020	10.1109/TTS.2020.2974991
Pickering J.E., Podsiadly M., Burnham K.J.	A Model-to-Decision Approach for the Autonomous Vehicle (AV) Ethical Dilemma: AV Collision with a Barrier/Pedestrian(s)	2019	10.1016/j.ifacol.2019.08.080
Potapov, A.; Rodionov, S.	Universal empathy and ethical bias for artificial general intelligence	2014	10.1080/0952813X.2014.895112
Poulsen, A.; Burmeister, O.	Overcoming carer shortages with care robots: Dynamic value trade-offs in run-time	2019	10.3127/ajis.v23i0.1688
Poulsen, A.; Burmeister, O. K.; Tien, D.	Care Robot Transparency Isn't Enough for Trust	2018	10.1109/TENCONSpring.2018.8692047
Pournaras, E.; Pilgerstorfer, P.; Asikis, T.	Decentralized Collective Learning for Self-managed Sharing Economies	2018	10.1145/3277668
Przegalinska A., Ciechanowski L., Stroz A., Gloor P., Mazurek G.	In bot we trust: A new methodology of chatbot performance measures	2019	10.1016/j.bushor.2019.08.005
Pugnetti, C.; Schlapfer, R.	Customer Preferences and Implicit Tradeoffs in Accident Scenarios for Self-Driving Vehicle Algorithms	2018	10.3390/jrfm11020028

Putnam, D., Kovacova, M., Valaskova, K., Stehel, V.	The Algorithmic Governance of Smart Mobility: Regulatory Mechanisms for Driverless Vehicle Technologies and Networked Automated Transport Systems	2019	10.22381/CRLSJ11120193
Radovanović, S.; Petrović, A.; Delibašić, B.; Suknović, M.	Making hospital readmission classifier fair – What is the cost? Central European Conference on Information and Intelligent Systems	2019	
Raff, E.; Sylvester, J.; Mills, S.	Fair Forests: Regularized Tree Induction to Minimize Model Bias	2018	10.1145/3278721.3278742
Raji I.D., Buolamwini J.	Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products	2019	10.1145/3306618.3314244
Rhim J., Lee G.-B., Lee J.-H.	Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of Korea and Canada	2020	10.1016/j.chb.2019.08.010
Riegler C.	The moral decision-making capacity of self-driving cars: Socially responsible technological development, algorithm-driven sensing devices, and autonomous vehicle ethics	2019	10.22381/CRLSJ11120192
Rivas, P.; Chelsi, C.; Nishit, N.; Ravula, L.	Application-Agnostic Chatbot Deployment Considerations: A Case Study	2019	10.1109/CSCI49370.2019.00070
Rivas, P.; Holzmayer, K.; Hernandez, C.; Grippaldi, C.	Excitement and Concerns about Machine Learning-Based Chatbots and Talkbots: A Survey	2018	10.1109/ISTAS.2018.8638280
Robertson, L. J.; Abbas, R.; Alici, G.; Munoz, A.; Michael, K.	Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots	2019	10.1109/JPROC.2018.2889678
Rodrigo L. Cardoso, Wagner Meira Jr. , Virgilio Almeida, Mohammed J. Zaki	A Framework for Benchmarking Discrimination-Aware Models in Machine Learning	2019	10.1145/3306618.3314262
Rodriguez-Soto, M.; Lopez-Sanchez, M.; Rodriguez-Aguilar, J.A.	A Structural Solution to Sequential Moral Dilemmas	2020	10.5555/3398761.3398895
Rowthorn, M.	How Should Autonomous Vehicles Make Moral Decisions? Machine Ethics, Artificial Driving Intelligence, and Crash Algorithms, Contemporary Readings in Law and Social Justice	2019	10.22381/CRLSJ11120191.

Rubel A., Castro C., Pham A.	Agency Laundering and Information Technologies	2019	10.1007/s10677-019-10030-w
Ryan, M.; Antoniou, J.; Brooks, L.; Jiya, T.; Macnish, K.; Stahl, B.	Technofixing the Future: Ethical Side Effects of Using AI and Big Data to Meet the SDGs	2019	10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00101
Sahu, S.; Singh, S. K.	Ethics in AI: Collaborative filtering based approach to alleviate strong user biases and prejudices	2019	10.1109/IC3.2019.8844875
Saltz, J.; Skirpan, M.; Fiesler, C.; Gorelick, M.; Yeh, T.; Heckman, R.; Dewar, N.; Beard, N.	Integrating Ethics within Machine Learning Courses	2019	10.1145/3341164.
Saxena, N.A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D.C.; Liu, Y.	How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness	2019	10.1145/3306618.3314248
Schelter S.; He Y.; Khilnani J.; Stoyanovich J.	FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions	2020	10.5441/002/edbt.2020.41
Schuelke-Leech, B.; Leech, T. C.; Barry, B.; Jordan-Mattingly, S.	Ethical Dilemmas for Engineers in the Development of Autonomous Systems	2018	10.1109/ISTAS.2018.8638282
Sendak, M.; Elish, M.C.; Gao, M.; Futoma, J.; Ratliff, W.; Nichols, M.; Bedoya, A.; Balu, S; O’Brien, C.	“The human body is a black box”: supporting clinical decision-making with deep learning	2020	10.1145/3351095.3372827
Senft E., Lemaignan S., Baxter P.E., Bartlett M., Belpaeme T.	Teaching robots social autonomy from in situ human guidance	2019	10.1126/scirobotics.aat1186
Serramia, M.; Lopez-Sanchez, M.; Rodriguez-Aguilar, J.A.; Rodriguez, M.; Wooldridge, M.; Morales, J.; Ansotegui, C.	Moral Values in Norm Decision Making	2018	10.5555/3237383.3237891
Shank D.B., DeSanti A.	Attributions of morality and mind to artificial intelligence after real-world moral violations	2018	10.1016/j.chb.2018.05.014
Shank D.B., Graves C., Gott A., Gamez P., Rodriguez S.	Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence	2019	10.1016/j.chb.2019.04.001
Sharma, S.; Henderson, J.; Ghosh, J.	CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models	2020	10.1145/3375627.3375812

Sharma,S.; Zhang, Y.; Ríos Aliaga, J.M.; Bouneffouf, D.; Muthusamy, V.; Varshney, K.R.	Data Augmentation for Discrimination Prevention and Bias Disambiguation	2020	10.1145/3375627.3375865
Shi, W.; Wang, X.; Oh, Y.J.; Zhang, J.; Sahay, S.; Yu, Z.	Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies	2020	10.1145/3313831.3376843
Sholla, S.; Mir, R. N.; Chishti, M. A.	Towards the design of ethics aware systems for the Internet of Things	2019	10.23919/JCC.2019.09.016
Shulman, E.; Wolf, L.	Meta Decision Trees for Explainable Recommendation Systems	2020	10.1145/3375627.3375876
Shvo, M.	Towards Empathetic Planning and Plan Recognition	2019	10.1145/3306618.3314307
Sivill,T.	Ethical and Statistical Considerations in Models of Moral Judgments	2019	10.3389/frobt.2019.00039
Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H.	Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods	2020	10.1145/3375627.3375830
Smith B.	Personality facets and ethics positions as directives for self-driving vehicles	2019	10.1016/j.techsoc.2018.12.006
Srinivasan R., Chander A.	Understanding bias in datasets using topological data analysis	2019	
Srivastava, B.; Rossi, F.	Towards Composable Bias Rating of AI Services	2018	10.1145/3278721.3278744
Stock, O., Guerini, M., Pianesi, F.	Ethical dilemmas for adaptive persuasion systems	2016	
Stowers, K.; Leyva, K.; Hancock, G. M.; Hancock, Peter A.	Life or Death by Robot?	2016	10.1177/1064804616635811
Suárez-Gonzalo S., Mas-Manchón L., Guerrero-Solé F.	Tay is You. The attribution of responsibility in the algorithmic culture.	2019	10.15847/obsOBS13220191432
Suwa S., Tsujimura M., Ide H., Kodate N., Ishimaru M., Shimamura A., Yu W.	Home-care Professionals’ Ethical Perceptions of the Development and Use of Home-care Robots for Older Adults in Japan	2020	10.1080/10447318.2020.1736809
Teso, S.; Kersting, K.	Explanatory Interactive Machine Learning	2019	10.1145/3306618.3314293
Tho, Q. H.; Phap, H. C.; Phuong, P. A.	A solution to ethical and legal problem with the decision-making model of autonomous vehicles	2019	10.1109/KSE.2019.8919452
Thornton, S.M.; Pan, S.; Erlien, S.M.; Gerdes, J. C.	Incorporating Ethical Considerations Into Automated Vehicle Control	2017	10.1109/TITS.2016.2609339

Trentin, V.; da Silva Guerra, R.; Rubert Librelotto, G.	Contradictions in Assessing Human Morals and the Ethical Design of Autonomous Vehicles	2019	10.1109/LARS-SBR-WRE48964.2019.00072
Ullman, D.; Malle, B.F.	What Does it Mean to Trust a Robot? Steps Toward a Multi-dimensional Measure of Trust	2018	10.1145/3173386.3176991
Urquhart L., Reedman-Flint D., Leesakul N.	Responsible domestic robotics: exploring ethical implications of robots in the home	2019	10.1108/JICES-12-2018-0096
Ustun B., Liu Y., Parkes D.C.	Fairness without harm: Decoupled classifiers with preference guarantees	2019	
Vakkuri V., Kemell K.-K.	Implementing AI ethics in practice: An empirical evaluation of the RESOLVEDD strategy	2019	10.1007/978-3-030-33742-1_21
Vakkuri V., Kemell K.-K., Abrahamsson P.	Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study	2019	10.1007/978-3-030-35333-9_24
Vakkuri V., Kemell K., Kultanen J., Abrahamsson P.	The Current State of Industrial Practice in Artificial Intelligence Ethics	2020	10.1109/MS.2020.2985621
Vakkuri, V. ; Kemell, K.; Abrahamsson, P.	Ethically Aligned Design: An Empirical Evaluation of the RESOLVEDD-Strategy in Software and Systems Development Context	2019	10.1109/SEAA.2019.00015
Valles-Peris, N.; Angulo, C.; Domenech, M.	Children’s Imaginaries of Human-Robot Interaction in Healthcare	2018	10.3390/ijerph15050970.
Van Dang, C.; Jun,M.; Shin,Y-B; Choi,J-W; Kim,J-W.	Application of modified Asimov's laws to the agent of home service robot using state, operator, and result (Soar)	2018	10.1177/1729881418780822
van Kemenade M.A.M., Hoorn J.F., Konijn E.A.	Do you care for robots that care? Exploring the opinions of vocational care students on the use of healthcare robots	2019	10.3390/robotics8010022
Van Kemenade, M.A.M., Konijn, E.A., Hoorn, J.F.	Robots humanize care: Moral concerns versus witnessed benefits for the elderly	2015	10.5220/0005287706480653
van Maris A., Zook N., Caleb-Solly P., Studley M., Winfield A., Dogramadzi S.	Designing Ethical Social Robots—A Longitudinal Field Study With Older Adults	2020	10.3389/frobt.2020.00001
van Maris, A.; Sutherland, A.; Mazel, A.; Dogramadzi, S.; Zook, N.; Studley, M.; Winfield, A.; Caleb-Solly, P.	The Impact of Affective Verbal Expressions in Social Robots	2020	10.1145/3371382.3378358

Vanderelst D., Willems J.	Can We Agree on What Robots Should be Allowed to Do? An Exercise in Rule Selection for Ethical Care Robots	2019	10.1007/s12369-019-00612-0
Vanderelst D., Winfield A.	An architecture for ethical robots inspired by the simulation theory of cognition	2018	10.1016/j.cogsys.2017.04.002
Vanderelst, D.; Winfield, A.;	The Dark Side of Ethical Robots	2018	10.1145/3278721.3278726
Veale, M.; Van Kleek, M.; Binns, R.	Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making	2018	10.1145/3173574.3174014
Verdiesen, I.; Santoni de Sio, F. ; Dignum, V.	Moral Values Related to Autonomous Weapon Systems: An Empirical Survey that Reveals Common Ground for the Ethical Debate	2019	10.1109/MTS.2019.2948439
Vetrò A., Santangelo A., Beretta E., De Martin J.C.	AI: from rational agents to socially responsible agents	2019	10.1108/DPRG-08-2018-0049
Vrščaj D., Nyholm S., Verbong G.P.J.	Is tomorrow's car appealing today? Ethical issues and user attitudes beyond automation	2020	10.1007/s00146-020-00941-z
Wächter, L.; Lindner, F.	An Explorative Comparison of Blame Attributions to Companion Robots Across Various Moral Dilemmas	2018	10.1145/3284432.3284463
Wangmo T., Lipps M., Kressig R.W., Ienca M.	Ethical concerns with the use of intelligent assistive technology: Findings from a qualitative study with professional stakeholders	2019	10.1186/s12910-019-0437-z
Wasilow, S.; Thorpe, J. B.	Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective, AI Magazine	2019	
Webb, H., Patel, M., Rovatsos, M., Davoust, A., Ceppi, S., Koene, A., Dowthwaite, L., Portillo, V., Jirotka, M. and Cano, M.	"It would be pretty immoral to choose a random algorithm": Opening up algorithmic interpretability and transparency	2019	10.1108/JICES-11-2018-0092.
Wolf, L.; Galanti, T.; Hazan, T.	A Formal Approach to Explainability	2019	10.1145/3306618.3314260
Wouters N., Kelly R., Velloso E., Wolf K., Ferdous H.S., Newn J., Joukhadar Z., Vetere F.	Biometric mirror: Exploring values and attitudes towards facial analysis and automated decision-making	2019	10.1145/3322276.3322304

Wu Y.-H., Lin S.-D.	A low-cost ethics shaping approach for designing reinforcement learning agents	2018	
Yang, K.; Stoyanovich, J.; Asudeh, A.; Howe, B.; Jagadish, HV; Miklau, G.	A Nutritional Label for Rankings	2018	10.1145/3183713.3193568
Yilmaz, L.; Sivaraj, S.	A Cognitive Architecture for Verifiable System Ethics via Explainable Autonomy	2019	10.1109/SYSCON.2019.8836896
Young A.D., Monroe A.E.	Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas	2019	10.1016/j.jesp.2019.103870
Yu, H.; Liu, Z.; Liu, Y.; Chen, T.; Cong, M.; Weng, X.; Niyato, D.; Yang, Q.	A Fairness-aware Incentive Scheme for Federated Learning	2020	10.1145/3375627.3375840
Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F.	Systematic review of research on artificial intelligence applications in higher education – where are the educators?: Revista de Universidad y Sociedad del Conocimiento	2019	10.1186/s41239-019-0171-0.
Zhang, B.; Dafoe, A.	U.S. Public Opinion on the Governance of Artificial Intelligence	2020	10.1145/3375627.3375827
Zhou, J.; Chen, F.	DecisionMind: revealing human cognition states in data analytics-driven decision making with a multimodal interface	2018	10.1007/s12193-017-0249-8