Information & Management 61 (2024) 103969

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders

Joakim Laine, Matti Minkkinen, Matti Mäntymäki [*]

Turku School of Economics, University of Turku, Finland

ARTICLE INFO

Keywords:
Artificial intelligence
Auditing
AI ethics
AI governance
AI auditing
Ethics-based AI auditing
Systematic literature review

ABSTRACT

This systematic literature review synthesizes the conceptualizations of ethical principles in AI auditing literature and the knowledge contributions to the stakeholders of AI auditing. We explain how the literature discusses fairness, transparency, non-maleficence, responsibility, privacy, trust, beneficence, and freedom/autonomy. Conceptualizations vary along social/technical- and process/outcome-oriented dimensions. The main stakeholders of ethics-based AI auditing are system developers and deployers, the wider public, researchers, auditors, AI system users, and regulators. AI auditing provides three types of knowledge contributions to stakeholders: 1) guidance; 2) methods, tools, and frameworks; and 3) awareness and empowerment.

1. Introduction

Artificial intelligence (AI) is a rapidly advancing research field comprising a set of technological capabilities that provide new resources, opportunities, and applications for organizations and society [12,28,91]. Although AI has numerous definitions, it generally refers to an information system's ability to interpret and learn from data and achieve goals through adaptation [61] and, more broadly, to an advancing frontier of computing [12]. Companies, public organizations, and governments are looking into AI to enable efficiency gains and new products and services (AI [1,11]). However, the rapid diffusion of AI systems has elicited risks concerning algorithmic opacity, unethical conduct, and unintended consequences, e.g., racial and gender biases [69,118], often referred to collectively as the 'dark side' of AI [90]. Alongside the concepts of trustworthy AI [63,81] and AI governance [13,92,105], the auditing of AI systems has been put forth as one response to these challenges [16,102,128]. Scholars, organizations, and decision-makers increasingly recognize the importance of auditing AI systems to ensure their socially acceptable and beneficial use [67,119]. This entails setting ethical, technical, social, and legal requirements to hold algorithms and data accountable to standards while establishing ethical principles, acceptable practices, and legislation [76].

Although AI auditing has been lauded as a "new industry" with great potential [67], it is currently in a formative stage (e.g., [93,107]). However, academic and gray literature on AI auditing is growing rapidly. AI auditing is a subfield of IT auditing (cf. [35]) that typically focuses on risk factors and control mechanisms. The Institute of Internal Auditors [56] defines internal auditing as "an independent, objective assurance and consulting activity designed to add value and improve an organization's operations" and states that it "helps an organization accomplish its objectives by bringing a systematic, disciplined approach to evaluate and improve the effectiveness of risk management, control, and governance processes." Raji et al. [118] cite the Institute for Electrical and Electronics Engineers (IEEE) definition of audit: "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures." Hence, an IT audit typically collects evidence on risks and control mechanisms and assesses an organization's technological infrastructure at a point in time to ensure the accuracy, efficiency, security, and compliance of systems and processes [35].

AI auditing encompasses internal and external auditing [113]. Internal auditing methods use internal system information, while external auditing typically relies only on publicly available information and model outputs and may not include intermediate models or training data [7,86]. Internal auditing often complements external auditing and evaluates how well a product or software fits expected system behavior standards [119]. In external auditing, the auditor is often either an independent auditor, an external auditor a company hires, or an auditor an

[*] Corresponding author at: FI-20014 University of Turku.
E-mail address: matti.mantymaki@utu.fi (M. Mäntymäki).

https://doi.org/10.1016/j.im.2024.103969
Received 17 December 2022; Received in revised form 25 April 2024; Accepted 30 April 2024
Available online 1 May 2024

outside source hires. The key elements for the auditor, in this case, are the creation of the algorithm, how the algorithm works, what data it needs, and the algorithm's expected outputs [83]. Both internal and external auditing are crucial parts of AI auditing, and neither should be excluded from comprehensive investigations a priori.

The AI auditing literature is fragmented, examining specific contexts such as search engines [21], facial recognition [73], social networks [144], e-commerce [149], and online job boards [55]. Thus far, few syntheses can be found in the literature, with an exception being Batarseh et al. [9], who scored AI assurance methods based on their applicability but devoted little attention to ethics issues. In its current formative state [93], the AI auditing field would benefit from clearly articulated and practical, implementable principles of AI ethics and, eventually, standards. Nevertheless, as various stakeholders understand problems differently, establishing standards is a significant challenge for the emerging AI auditing industry [23].

AI ethics principles provide a starting point for consolidating ethics-based AI auditing [106] and identifying ethical standards against which systems and organizations can be audited (cf. [113]). Thus far, companies, public organizations, and researchers have produced numerous lists of AI ethics principles (e.g., [59]). However, abstract principles require significant translation work to be implemented practically in AI systems [104,105,131]. To understand how ethics can inform AI auditing, it is essential to examine how AI ethics principles are articulated in the AI auditing literature and how they guide ethics-based AI auditing.

In addition to clarifying the conceptualizations of key principles, consolidating the AI auditing literature requires understanding its knowledge contributions to stakeholders. Thus far, little academic attention has been paid to stakeholders of AI auditing apart from initial sketches on actor networks ([139], p. 86; [94]). However, literature has been proliferating more broadly on algorithmic accountability as a networked phenomenon involving numerous stakeholders [89,94,148]. The stakeholder perspective is essential because AI development is often a multi-actor activity [74], AI systems are used in a multitude of sectors, and AI ethics issues concern stakeholder groups such as minorities.

To address these gaps and synthesize the fragmented AI auditing literature, we have conducted a systematic literature review (SLR) to understand ethical principles and stakeholders in ethics-based AI auditing. To accomplish this, this paper seeks to answer two research questions:

(1) How does the academic ethics-based AI auditing literature conceptualize AI ethics principles?
(2) What knowledge contributions does the literature present to the stakeholders of ethics-based AI auditing?

As the research questions indicate, this paper aims to summarize and make sense of the emerging AI auditing literature and unpack the knowledge contributions to different stakeholders. To keep the scope manageable, we utilize existing frameworks of AI ethics principles (AI [1,59]) and investigate how these principles are reflected in the literature, including how focal concepts are theorized, rather than develop new sets of principles or new theoretical propositions about AI auditing. We limit our scope to the discussion of AI ethics principles and stakeholders in the ethics-based AI auditing literature, leaving subsequent questions about the practical implementation of auditing procedures to future research.

By analyzing conceptualizations of focal concepts and identifying stakeholders and knowledge contributions in the AI auditing literature, this review makes two contributions to the IS literature on AI in organizations (e.g., [11,12,79]) and to the IS literature on AI system stakeholders (e.g., [75,87]). First, we extend the IS literature on AI in organizations by surveying the full spectrum of ethical principles of AI auditing and introducing a matrix that outlines the goals and conceptualizations of AI auditing along social/technical- and process/outcome-oriented approaches. Second, we contribute to the literature on AI system stakeholders (e.g., [75,87]) by presenting a framework that clarifies the knowledge contributions of AI auditing literature to stakeholders and, thereby, illuminating how AI auditing can benefit the multi-stakeholder ecosystem involved in AI management (e.g., [139]). Collectively, these contributions aim to consolidate the disconnected literature streams relevant to ethics-based AI auditing, ultimately supporting the development of AI management and auditing practices.

The remainder of the paper is structured as follows. First, Section 2 briefly defines ethics-based AI auditing and introduces frameworks for AI ethics principles. Section 3 presents the SLR methodology. Section 4 presents the findings concerning the ethical principles and the stakeholders of ethics-based AI auditing. Section 5 discusses conceptualizations of focal concepts, summarizes the stakeholders and knowledge contributions, the implications for IS research and practice, and the study's limitations, concluding the paper with a future research agenda.

2. Conceptual background: AI auditing and ethical principles

This study focuses on the type of AI auditing that aims to ensure that AI is ethical when evaluated against established ethical principles. The approach has been termed "ethical algorithm auditing" [16] and "ethics-based AI auditing" [102]. Ethics-based auditing is not the only approach to AI auditing – alternative approaches include, e.g., ensuring legality, accuracy, efficiency, or safety [39,57]. However, ethics-based auditing is a crucial subfield of AI auditing because AI ethics discussions among academia and practitioners continue to grow, laying the foundation for AI auditing and regulatory developments such as the European Union's proposed AI Act (AI [1,38,102]). Furthermore, translating ethical principles into practices has been identified as one of the critical problems in ensuring socially responsible AI systems [104]. At the same time, formal standards and requirements for AI auditing remain unestablished, leading to concerns over whitewashing companies' reputations by publishing abstract ethical principles and guidelines (e.g., [138]).

Ethics-based AI auditing can be defined from a consequentialist (i.e., focusing on consequences) or deontological (i.e., focusing on duties and adherence to rules) perspective (cf. [48]). Brown et al. [16] focus on consequences and define ethical algorithm audits as "assessments of the algorithm's negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that gives rise to these negative impacts." Mökander et al. [102], in turn, opt for a deontological perspective and define ethics-based auditing of automated decision-making systems as "a structured process whereby an entity's present or past behavior is assessed for consistency with relevant principles or norms." Brown et al. [16] start with negative impacts on stakeholders, while Mökander et al. [102] highlight consistency with principles and norms.

In alignment with the research questions, this paper adopts the deontological perspective [102] and investigates how the emerging AI auditing literature conceptualizes ethical principles. At present, the ethics-based AI auditing literature is nascent and includes definitions and scoping of ethics-based auditing [102], discussions of best practices [103,107], and tentative auditing frameworks for algorithms [16,76,119]. Beyond the emerging ethics-based auditing core literature, the algorithmic auditing literature tends to focus on particular issues, such as bias, and use cases, such as search engines [21], facial recognition [73], and social networks [144]. More indirectly, discussions on AI ethics principles [48,59] also feed into the ethics-based AI auditing discussion.

The central issue for deontological ethics-based AI auditing is the adherence of AI systems to AI ethics principles, leading to repeated calls to get from principles to practice [59,102]. Thus, to differentiate ethics-based AI auditing from other types of auditing and to identify the relevant literature, we need an established framework of such principles. Ethics in the context of AI is a vast area with longstanding discussions, as evidenced by the fact that the Stanford Encyclopedia of Philosophy lists ten central debates on the ethics of AI, including bias, opacity, and manipulation of behavior ([99]; cf. [48]). In this paper, we adopt a pragmatic stance toward defining ethics; therefore, we omit conceptualizing ethics in the abstract and focus the ethics perspective on core AI ethics principles, which is common practice in the AI ethics literature (e.g., [40,41,48,59]).

To keep the review's scope manageable and in line with our research questions, we focus specifically on ethical principles articulated from the AI ethics and AI auditing literature. The AI ethics principles literature is extensive, and several overviews of ethics principles have been published (e.g., [48,59]). The EU-appointed High-Level Expert Group on Artificial Intelligence's *Ethical Guidelines for Trustworthy AI* (AI [1]) presents a credible source of synthesized expert knowledge. The EU's High-Level Expert Group's (AI HLEG) is an independent group set up by the European Commission to propose guidelines for trustworthy AI. Its guideline document (AI [1]) provides a suitable AI ethics framework for two reasons: It comes from an institutionally credible source due to the close EU connection and the presence of prominent academics and practitioners, and the framework is cited frequently and is known widely in the AI industry and beyond. The starting point for the expert group's work was to outline an approach to AI ethics aligned with fundamental rights, democracy, and the rule of law (AI [1]). The AI HLEG [1] puts forth four core ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. These principles are aligned with other discussions of fundamental ethics principles (e.g., [41,59]). Thus, they provide an appropriate entry point to introduce the principles briefly.

Respect for human autonomy refers to the principle that AI systems should complement rather than subordinate or manipulate humans and that humans must maintain full self-determination (AI [1]). Autonomy can be conceptualized in terms of positive freedoms (i.e., human flourishing and self-determination) or negative freedoms (i.e., freedom from manipulation and surveillance) [59]. Autonomy is also about consciously striking a balance between the decision-making power retained for humans and that delegated to artificial agents, thereby maintaining a desired level of human autonomy [41].

Prevention of harm denotes that AI systems should not cause or exacerbate mental or physical harm and should not be open to malicious use (AI [1]). The AI ethics guidelines literature features calls for safety, security, and prevention of foreseeable and unintentional harm [59]. Threats of an AI arms race and threats to privacy are repeatedly taken up [41].

Fairness means that AI systems respect both substantive fairness (freedom from bias and discrimination and equal opportunity) and procedural fairness (the ability to contest and seek redress) (AI [1]). Fairness is linked to justice in terms of outcomes and processes and to risks of bias in AI systems [41]. Fairness also covers diversity, inclusion, and equality issues pertinent to the labor market and across society [59].

Explicability refers to the principle that AI systems should be transparent about their capabilities and purposes and that their decisions should be explainable to those affected to the extent possible (AI [1]; cf. [75]). Calls for transparency cover data use, human-AI interaction, automated decisions, and the purpose of data use [59]. Moreover, explicability has been seen as complementing the other principles because, for AI systems to respect autonomy, prevention of harm, and fairness, we need to have an adequate understanding of their actions [41].

Jobin et al. [59] elaborated on these core principles by empirically analyzing the corpus of AI ethics principles and guidelines (84 documents) and revealing how certain principles came up more than others, how they were linked together, why they are important, what actors they pertain to, and how they should be implemented. Their research was conducted as a scoping review using academic literature. The study revealed five ethical principles that drive the current ethical AI discussion: transparency, justice and fairness, non-maleficence, responsibility, and privacy. Furthermore, they identified six other ethical principles that are less prevalent [59]. As the study's themes corresponded strongly with this SLR's scope, these existing principles were taken as the baseline to investigate how principles are conceptualized in the ethics-based AI auditing literature and how they relate to one another. In the subsequent methodology section, we elaborate on our use of both the AI HLEG [1] principles and the principles found in Jobin et al.'s [59] analysis.

3. Methodology

The reporting strategy in our SLR follows the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines and the PRISMA 2009 checklist [97]. The PRISMA framework was created to 1) ensure comprehensive reporting in the review, 2) provide enough methodological details, 3) reduce author bias in the results, 4) consider quality differences between studies, and 5) avoid misinterpretation or inadvertent bias in the results [130]. This section goes through the three main parts of the SLR process: literature search, data extraction, and data analysis.

3.1. Literature search

To find all the relevant literature for answering our research questions, we formulated a search syntax as well as inclusion and exclusion criteria. The search syntax was established using the following keywords: "auditing"; "artificial intelligence"; "AI"; "deep learning"; "machine learning"; "black box"; "algorithmic"; and "algorithm. The search syntax had to include the term "auditing" and at least one other keyword. Table 1 presents the search syntax.

Although ethics is not explicitly included as a search keyword, it is important to note that these keywords were chosen based on synonyms or close concepts of artificial intelligence to ensure that we captured all relevant literature on AI auditing, regardless of whether they explicitly mentioned ethics. In a subsequent phase, we manually identified papers focusing on ethical considerations in AI auditing.

The inclusion criteria in the literature search dictated that publications had to be published in peer-reviewed journals or conference proceedings (IC#1) in English (IC#2). Books, book chapters, and reviews were excluded (EC#1). In addition, we excluded studies in languages other than English (EC#2). We included studies that address the auditing of AI (IC#3), meaning that they discuss, for example, concepts and methods relevant to auditors. However, the studies did not need to discuss the entire AI auditing process; instead, they may present tools for auditing specific aspects, such as bias. We included studies published before March 2022 (IC#4). The initial search recognized all the auditing studies, and during the final step, we divided the papers into ethics-based and non-ethics-based studies. Table 2 presents the inclusion and exclusion criteria.

We used the following databases in the literature search: 1) Scopus; 2) Web of Science Core Collection; 3) IEEE Xplore; and 4) ACM Digital Library (conference proceedings and journal publications). Consulting several databases during an SLR is useful because it increases the comprehensiveness of the search, enables the identification of relevant studies that may be missed in a narrower search, and helps minimize bias in the selection of studies (e.g., [15]). We found including several databases beneficial because new search results surfaced in each database.

Database search started with Scopus, yielding a total of 672 studies. With Scopus, the search included only titles, abstracts, and keywords. With the same search, the Web of Science Core Collection database yielded 222 studies. We conducted full-text searches in the IEEE Xplore and ACM Digital Library databases. In the other databases, a full-text search was not possible. IEEE yielded 666 studies, ACM conferences 1485, and ACM journals 286. These searches included EC1 and EC2. The

Table 1
The Search syntax.

"auditing" AND ("artificial intelligence" OR "AI" OR "deep learning" OR "machine learning" OR "black box" OR "algorithmic" OR "algorithm")

Table 2
Inclusion and exclusion criteria.

Inclusion criteria (IC)		Exclusion criteria (EC)	
IC#1	Journal articles or conference papers only	**EC#1**	Books, book chapters, reviews, etc.
IC#2	Studies published in the English language	**EC#2**	Studies in languages other than English
IC#3	Studies that address the auditing of artificial intelligence	**EC#3**	Studies that focus on something other than the auditing of artificial intelligence, e.g., the use of artificial intelligence in auditing
IC#4	Studies published before March 2022		

total number of studies identified from the database search was 3331.

During the second phase, we screened the studies based on the inclusion and exclusion criteria outlined in Table 2. This screening process involved reading each article's title, abstract, and keywords. Additionally, we identified and removed any potential duplicate studies. Each article that appeared to fulfill the scope of our study and met our predefined inclusion criteria progressed to the next phase of our selection process, which was a full-text review. Of the 3331 articles retrieved during the first phase, 2988 were excluded in the second phase, leaving us with a sample of 343 studies that met our inclusion criteria.

The third phase entailed screening the 343 studies based on reading the full texts and reapplying the inclusion and exclusion criteria. At this stage, we excluded 233 studies, leaving us with 110 studies. Table 3 presents the number of studies after the database search and after the two phases of screening.

From the sample of 110 studies, we conducted backward citation chaining. A backward citation search aims to find papers that influenced the reviewed studies and, thus, ensure that no relevant studies are missing [53]. We screened the reference lists of the 110 studies based on their titles, included potentially relevant studies, and subjected them to a full-text screening and application of the inclusion and exclusion criteria. As a result of the backward citation chaining process, we augmented the sample with 30 additional studies, yielding 140 studies.

To identify the studies focusing particularly on ethics-based auditing, we searched the 140 studies for the AI HLEG ethical principles and requirements [1]. We focused on the ethical component of trustworthy AI, which the AI HLEG [1] divides into four principles: respect for human autonomy, prevention of harm, fairness, and explicability. These outlined principles are also translated into seven more concrete requirements: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness, 6) societal and environmental wellbeing, and 7) accountability.

Table 3
Number of studies per database.

Data source	Phase 1: database search	Phase 2: screening based on titles, abstracts, and keywords	Phase 3: screening based on full text
Scopus	672	116	51
Web of Science Core collection	222	12	4
IEEE Xplore	666	67	11
ACM Digital Library conference proceedings	1485	116	33
ACM Digital Library journal publications	286	32	11
Total number of studies	3331	343	110

After carefully examining each paper, we classified them into either ethics-based or non-ethics-based categories based on whether they explicitly discussed at least one of the AI HLEG ethical principles and requirements. Single mentions of ethical principles were not counted as a discussion. Specifically, we looked for evidence of discussing the four ethical principles of respect for human autonomy, prevention of harm, fairness, and explicability, as well as the seven concrete requirements outlined by AI HLEG. Only studies demonstrating a sustained discussion of these ethical principles and requirements were included in the ethics-based category, while the remaining studies were classified as non-ethics-based. For clarity, being a non-ethics-based paper does not mean the paper completely neglects ethics but that the core focus is elsewhere, e.g., on technical aspects. We divided the 140 studies into ethics-based ($n = 93$) and non-ethics-based ($n = 47$) studies. As a result, the final sample comprised 93 studies on ethics-based auditing of AI.

3.2. *Data extraction and analysis*

We analyzed the 93 studies in three steps. First, we collected descriptive data from every paper, including publication years, research methods, and the studies' main outputs. This helped form an overall understanding of the ethics-based AI auditing literature.

Second, we focused on analyzing the studies to uncover the conceptualizations of the ethical AI principles that Jobin et al. [59] summarized. The purpose of this analysis stage was to elicit how ethical AI principles are conceptualized in the ethics-based AI auditing literature, the most prevalent principles, and the linkages between the principles.

Jobin et al.'s [59] framework on global ethical AI guidelines was selected to guide our data analysis because we needed an elaborate framework for analyzing how the studies conceptualize ethical principles after the initial division into ethics-based and non-ethics-based studies. The Jobin et al. framework is used in addition to the previously discussed AI HLEG [1] for two reasons. First, a second framework is beneficial in reducing the risk of circular reasoning, i.e., selecting studies based on criteria from the AI HLEG and then analyzing them using the same criteria (cf. [33]). Second, as a scholarly summary of ethical principles, the Jobin et al. framework provides tools for qualitative analysis more readily than simpler policy-oriented lists. After consulting several overviews of AI ethics principles (e.g., [41,48]), we decided on Jobin et al.'s [59] highly cited categorization.

As a result, Based on Jobin et al. [59], we established a categorization of the ethical principles discussed in the studies. Each category included several lower-level principles and was analyzed separately to determine how the principles are conceptualized in AI auditing literature.

Finally, we analyzed the 93 studies to identify the knowledge contributions to ethics-based AI auditing stakeholder groups. The stakeholder groups in the reviewed studies were system developers and deployers, the wider public, researchers, auditors, AI system users, and regulators. Fig. 1 summarizes the entire SLR process.

4. Findings

In this section, we report the findings in the same order that the analysis was conducted: descriptive details, ethical principles, and stakeholders.

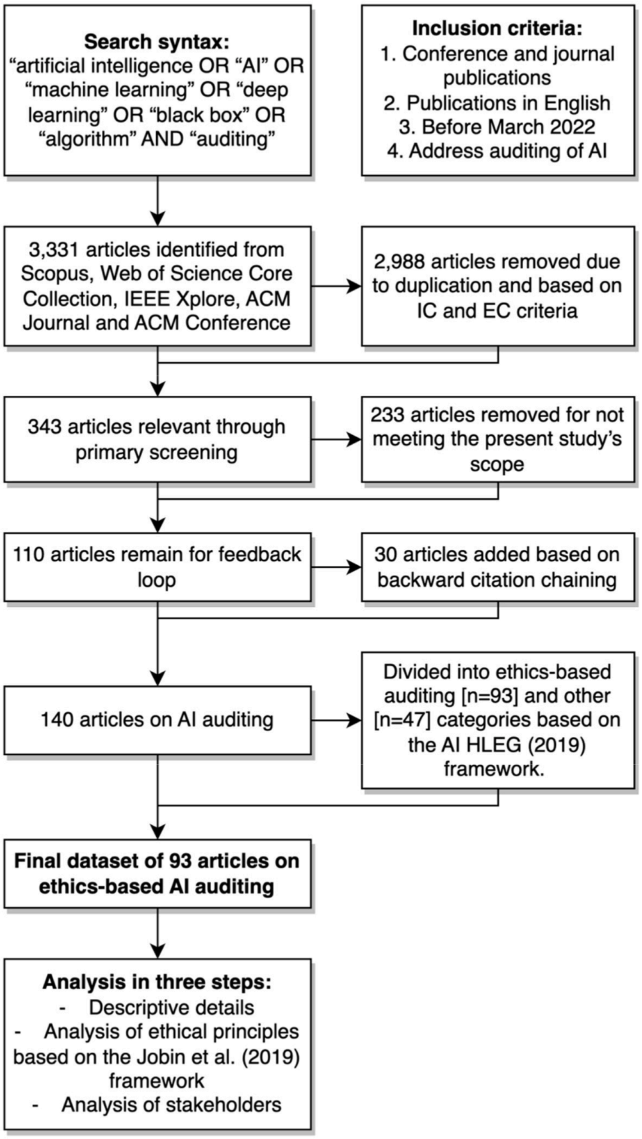
Search syntax:
“artificial intelligence OR “AI” OR “machine learning” OR “deep learning” OR “black box” OR “algorithm” AND “auditing”
Inclusion criteria:
1. Conference and journal publications
2. Publications in English
3. Before March 2022
4. Address auditing of AI
3,331 articles identified from Scopus, Web of Science Core Collection, IEEE Xplore, ACM Journal and ACM Conference
2,988 articles removed due to duplication and based on IC and EC criteria
343 articles relevant through primary screening
233 articles removed for not meeting the present study’s scope
110 articles remain for feedback loop
30 articles added based on backward citation chaining
140 articles on AI auditing
Divided into ethics-based auditing [n=93] and other [n=47] categories based on the AI HLEG (2019) framework.
Final dataset of 93 articles on ethics-based AI auditing
Analysis in three steps:
- Descriptive details
- Analysis of ethical principles based on the Jobin et al. (2019) framework
- Analysis of stakeholders

Fig. 1. The SLR process.

4.1. Descriptive details

The full sample from the databases and citation chaining is provided in Appendix A, highlighting each study’s research method and main outputs. Studies P29–P35 and P43–P50 were identified through citation chaining; the other studies came from the initial search.

The methods were derived from explicit statements made in the study, or when these were not available, using our best judgment. Some studies utilized more than one method, and these studies were classed under all the relevant methods. An empirical approach, which included quantitative, qualitative, experiment, design science, and case studies, was adopted by 68 studies, while 30 studies were conceptual. Among the empirical studies, the design science approach was the most common (29 studies). Although these findings are indicative rather than conclusive, they suggest that concrete empirical studies and tools are appreciated in the ethics-based AI auditing literature. Appendix B shows the ethical principles and methods in the reviewed studies.

4.2. Ethical principles in AI auditing

As was established previously, the field of ethics-based AI auditing is relatively new; therefore, examining how ethical principles are currently discussed and utilized is essential to understanding the emerging field. This section focuses on conceptualizations of ethical principles in the ethics-based AI auditing literature, that is, how frequently the ethical principles identified in existing AI guidelines by Jobin et al. [59] appear and how they are conceptualized.

Table 4 provides the total number of studies in the sample discussing a particular principle and the terms attached to each principle. The rows are the top-level principles that Jobin et al. [59] identified, and the included codes for each principle are indicated in the last column. If these terms were present in a paper, it was viewed as discussing the related principle. To avoid spurious matches, the term had to be linked with AI auditing to match the criteria. The three final principles (sustainability, dignity, and solidarity) are listed for comprehensiveness, even though none of the studies mentioned them.

The following sections analyze conceptualizations of these principles found in the literature on AI auditing. As we can see, justice and fairness—followed by transparency, non-maleficence, and responsibility—were the most common principles. The literature discussed privacy, trust, beneficence, and freedom/autonomy to a lesser extent (fewer than 30 of 93 studies).

4.2.1. Justice and fairness (84 studies)

Featured in 84 of our 93 sources, justice and fairness is by far the most prevalent principle category in the AI auditing literature. “Fairness” and “bias” were the most mentioned codes in this category. “Fairness” occurred in titles 20 times, while “bias” appeared 11 times. For clarity, in the following passage, “fairness” refers only to the specific term “fairness” rather than including all related terms. Overall, the word “fairness” was present in 68 sources.

Fairness is a multifaceted concept. Several of the reviewed AI auditing papers highlighted the multiple definitions of fairness and the concomitant challenges [10,19,42,47,50,63,69,77,109,111,150]. No single unified definition of fairness exists, and the reviewed studies articulated different perspectives on fairness.

Table 4
Number of studies and ethical principles.

	Total no. of studies	Included terms
Justice and fairness	83	Justice, fairness, consistency, inclusion, equality, equity, bias, discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access, distribution
Transparency	54	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Non-maleficence	41	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity, non-subversion
Responsibility	39	Responsibility, accountability, liability, acting with integrity
Privacy	25	Privacy, personal or private information
Trust	22	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment, trust
Beneficence	20	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	13	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Sustainability	(not discussed in the reviewed studies)	Sustainability, environment, energy, resources
Dignity	(not discussed in the reviewed studies)	Dignity
Solidarity	(not discussed in the reviewed studies)	Solidarity, social security, cohesion

For AI auditing purposes, the conceptualizations of fairness considered, e.g., addressing specific problems [49,50], different metrics that depend on context and culture [6,10,19,47,109,111,150], statistical definitions [64,116], and ethical principles that denote what are viewed as fair actions [63,77,95,141]. AI researcher Arvind Narayanan calls the attempt to find a single definition of fairness in computer science "a wild goose chase," describing at least 21 mathematical definitions of fairness from the literature [10,47]. Similarly, Barlas et al. [4] noted that fairness "is best understood as a placeholder term for a variety of normative egalitarian considerations."

Conceptualizations of bias are also widely represented in the AI auditing literature. Galdon Clavell et al. [42] defined bias as unfavorable treatment of an already disadvantaged group. Further, they stated that the criteria by which something constitutes bias should be framed from a social and ethical perspective [42]. The racial bias concept, associated with image tagging, for example, is one of the most common topics and is related closely to fairness [49,55,71,73,88]. From the bias perspective, fairness was conceptualized primarily as a human identity issue or an issue of a specific group or individual receiving unfavorable treatment [34,111,117,125,137,141]. Race and gender have become two major concerns regarding bias in the machine learning (ML) fairness literature, in which fairness then means treating subjects similarly regardless of their protected attributes, i.e., characteristics such as gender, that cannot be used as decision criteria [126,141].

Fairness in AI auditing can be conceptualized as a lack of bias because, in the decision-making process, fairness can be viewed as the absence of any preconceptions, discrimination, or favor toward an individual or group based on their inherent or acquired characteristics [85]. However, numerous definitions of fairness and bias have made discovering biases in ML models challenging. Naturally encoded societal biases in ML models are often called algorithmic biases. These types of biases should be addressed before deploying ML systems, which is why it is vital to audit ML models. In addition to multiple fairness definitions, recognizing biases can be challenging due to the inherent intersectionality of bias, i.e., populations are defined by multiple features [19].

Reflecting the complexity of the fairness concept, the AI auditing literature makes a distinction between group fairness and individual fairness [6,10,14,19,21,25,42,49,64,65,102,111,117,118,125]. Group fairness is defined as "the goal of groups defined by protected attributes (an attribute that partitions a population into groups that have parity in terms of benefit received) receiving similar treatments or outcomes" [10]. In contrast, individual fairness is "the goal of similar individuals receiving similar treatments or outcomes" [10]. Audits of algorithmic systems typically highlight the notion of group fairness, which holds that advantaged and protected groups should be treated the same as others. By comparison, individual fairness deals with consistency, i.e., the notion that systems should treat similar individuals the same [6].

For auditing purposes, fairness and bias can be addressed technically, but they are also socio-political issues that influence, e.g., public opinion [122]; thus, they are viewed as requiring more comprehensive algorithm audits. Technical bias detection mechanisms can help ensure, e.g., fairness in the decision-making processes of autonomous software [141], as well as detecting bias in language use [111].

4.2.2. Transparency (53 studies)

Transparency in the context of information ethics can be conceptualized as "the availability of information, the conditions of accessibility, and how the … information may pragmatically or epistemically support the user's decision-making process" (Turilli and Floridi, cited in [95]). More specifically for algorithmic models, Shulner-Tal et al. [134] defined transparency as "the degree to which the model is understandable by itself and mainly refers to the characteristics of the model." Transparency ensures that the AI system provides information about its decision-making processes and that various stakeholders understand the performance and limitations of the system impacting them by explaining, interpreting, and reproducing its decisions [63,98].

Conceptualizations of transparency exhibited four interlinked perspectives in the AI auditing studies. First, transparency was linked primarily with general data ethics principles, fairness, and accountability [5,37,45,47,62,63,70,82,93,100,115,116,118,134,135,147,149]. Second, transparency was viewed as operationalized via algorithmic auditing to lead to better technologies [6,14,42,47,77,118,121,124]. Third, transparency was viewed as a way to minimize harm and improve AI [5,16,18,32,58,62,73,115,117,120,123,142]. Finally, transparency was viewed as a way to improve responsibility and explainability issues [24,40,46,63,82,83,95,98,116,133–136,138]. In sum, transparency seems to be a focal principle that links to most other ethical principles.

Like fairness, transparency includes several dimensions. *Transparency of architecture* measures how well stakeholders know the algorithms' structure. *Transparency of use* measures how algorithms are used. In turn, *transparency of data and use* measures how much stakeholders know about the collection and subsequent use of data for the algorithm [16].

Conceptualizations in the AI auditing literature link transparency closely to accountability. Although accountability focuses on methods for holding a system to an ethical standard determined by domain experts, transparency refers to understanding the mechanisms behind why algorithms produce specific outputs [47]. The meaning of both transparency and accountability is to protect the user against undesirable or harmful results and to ensure the application of laws appropriate to digital environments [123]. However, transparency can be viewed as a broader concept than accountability because a system's property that provides visibility of its governing norms and behavior [121]. Loi and Spielkamp [82] argued that lack of transparency is the only reason for discussing accountability problems. They defined transparency as a way to make information accessible and an entitlement of a counterpart outside the accountable organization to obtain that information. Transparency as an internal control can be controlled via activities, e.g., documenting processes, monitoring relevant events, building backups and contingency plans, and effectively controlling data assets and ML algorithms [82].

Transparency is crucial for auditing because it exposes AI systems' inner workings and enables reporting on their operations. Inner mechanisms expose the critical knobs of the decision-making process, which later helps developers apply a code of ethics in ML systems [47]. A transparency report makes it possible to understand why the model behaves in a certain way, e.g., revealing reasons for the model's bias [14]. Transparency reports require reasons why and how ML technology makes decisions and is the most crucial accountability attribute for liability questions [121]. Without sufficient data pre-processing transparency, it is difficult to identify a potential threat in an otherwise effective model [147]. Different methods to make AI systems transparent and explainable can be divided into pre-modeling (explaining data sets), in-modeling (making interpretable models), and post-modeling (building proxy models) approaches [63].

Auditing processes are being developed to make algorithms more transparent for users and to promote fairness in an effort to "open the black box" [5,82,124,149]. Fairer and more transparent systems require that developers engage with social and legal facets of fairness, develop software that concretizes these values, and undergo an independent algorithm audit to ensure technical correctness and social accountability—but few companies match these criteria transparently [149]. According to the literature, external pressure is necessary to direct corporations toward transparency, accountability, and fairness because companies hesitate to disclose details about their systems [117]. Thus, professional associations, e.g., IEEE and ACM, encourage developers to take measures to promote transparency in the algorithmic systems they build [73]. Public scandals have emerged over the ethical impacts of AI systems, and lack of transparency and data misuse have often been critical issues in these scandals [16].

Transparency is tied conceptually to technical explainability and AI systems' trustworthiness. Both transparency and trustworthiness can improve users' perceptions of the system, including its fairness, and may affect their willingness to use it [134]. The literature presents two types of explainability of AI systems, enabling different AI auditing approaches. Whereas white-box explainability provides explanations for interpretable algorithms that reveal their structure, black-box explanations "refer to the ability to present a justified outcome to the user based on interpretable models" [134]. Methods that fall under the transparent-by-design category aim to train ML models that are both interpretable and accurate [109]. In systems where full transparency is impossible, algorithm auditing should happen externally. In this regard, people, organizations, or other audit targets tend to dismiss audit results if any dishonesty or non-transparency occurs in the audit methodology, so auditors need to live up to high ethical standards [100,118].

Transparency sets requirements for both systems and human stakeholders of AI auditing. For an ML system to be transparent, algorithmic tools must be open, and users and developers must have the skills to understand them [6,73]. A key limitation of many AI tools is a lack of transparency and explainability. For example, the AI Fairness 360 tool, an open-source toolkit to mitigate bias in ML models, is not transparent about how the input data were reweighed and distributed [116].

4.2.3. Non-maleficence (42 studies)

The reviewed studies discussing non-maleficence primarily state that AI should not cause any harm. A clear definition of non-maleficence did not appear in the literature. Codes mainly addressed other dimensions, e.g., fairness and bias issues [42,46,65,76,88,117–119,124,132,146], accountability and liability issues [24,63,82,83,135,136], security issues [2,24,151], or potential harms [16,26,58,63,85,93,96,121,133]. Thus, non-maleficence is conceptually fuzzy. Moreover, it is often equated with beneficence because beneficence is based on doing only good and non-maleficence is based on doing no harm [40,76,77]. Floridi et al. [40] identified a further link between non-maleficence and justice, defining justice as preventing the creation of new harms and ensuring that AI creates benefits and eliminates unfair discrimination.

One of the most important reasons for auditing is to identify and prevent harmful repercussions [100], which can include examining broad classes of systems, such as search, e-commerce, news recommendations, online advertising, maps, ridesharing, online reviews and ratings, natural language processing, and recommendations [149]. Several organizations have proposed methods and guidelines to prevent harm and make AI safe, reliable, and trustworthy [63]. However, it is crucial to separate system reliability harms from societal harms [63]. An AI system might be technically reliable but not meet ethical expectations [118]. Potential sources of harm and social impacts are then evaluated through auditing. Ethics-based auditing can help anticipate possible negative consequences in three stages: pre-processing (reviewing input data), in-processing (model selection), and post-processing (calibrating odds) [102]. Moreover, social impact assessments have been suggested to identify harms caused by AI systems [6,119]. According to the reviewed literature, a robust regulatory system and concrete methods must be developed because public trust in AI will arise only by guaranteeing that the public is protected from the harmful consequences of AI [3,66].

Ethical AI auditing provides external information about doing no harm, i.e., detecting and calling out potential biases, harms, or flaws [76]. Therefore, an important goal of auditing is to minimize AI's harmful biases. Metaxa et al. [88] highlighted the relation to harm in their conceptualization of bias as a skew that produces harm, including the subcategory of representational harm, i.e., when a system perpetuates or exaggerates social inequalities along identity lines.

As AI systems become widespread, external pressure to address harmful biases increases [117]. Marginalized populations need to be protected, and ethical guidelines, policies, and corporate practices are needed to ensure that evolving AI technology does not cause harm. The need for audit studies to diagnose harmful discrimination has also been highlighted by Sandvig et al. [124] and Goodman [46]. In particular, within companies with big data repositories—e.g., Facebook, YouTube, or Google—it is essential to investigate their algorithms' potentially harmful and discriminating consequences.

AI systems should be technically robust, perform as intended, do no damage to other systems and society, and recover from failure without harming users [63]. However, even though audits have made progress in detecting biases and harmful AI behavior, AI developers still struggle to detect and mitigate harmful biases in their systems due to their own cultural blind spots [132].

4.2.4. Responsibility (38 studies)

The responsibility principle is discussed in the AI auditing literature, often through accountability, but rarely defined (cf. [30]). Loi and Spielkamp [82] argued that any definition of accountability must include at least three elements: 1) responsibility for actions and choices; 2) answerability, which includes the capacity and willingness to reveal the reasons behind decisions to a selected counterpart, and such counterpart's entitlement to request that such reasons be revealed; and 3) sanctionability of the accountable party. The second element, answerability, clearly connects accountability to transparency. In turn, Reed et al. [121] followed the Accountability for Cloud project's definition of responsibility: "the property of an organization or individual in relation to an object, process, or system of being assigned to take action to be in compliance with the norms." Primarily, themes around responsibility and accountability include recommendations for more responsible AI development and use [147]. Desired outcomes for accountable AI include allowing for auditing and documentation to hold organizations accountable for their AI-based products and services and explaining and justifying the actions to different users with whom the system interacts [63,98].

The need for accountability comes from the various high-stakes applications of algorithmic systems; hence, it has become essential to audit these algorithmic models' design, development, and implementation [63]. One of the most cited AI accountability studies is Raji et al.'s [118] framework for internal algorithmic auditing to close the AI accountability gap. They defined accountability as "the state of being responsible or answerable for a system, its behavior, and its potential impacts" ([118], p. 34). They noted that algorithms are not moral or legal agents; therefore, algorithms cannot be held accountable, but organizations designing algorithms can be accountable through governance structures [118]. A complementary view is to conceptualize algorithmic accountability more technically, e.g., as a method for holding a system to an ethical standard determined by domain experts [47]. Systems then should be able to apportion responsibility and determine who owns what particular occurrence [135]. This way, as natural and legal persons (people and organizations) are accountable for their own actions, as well as the actions of machines and systems under their control. For users, it is difficult to judge who is accountable for the results because AI algorithms often are integrated into larger systems, further exacerbating opaqueness and ambiguity about ownership [147].

In the AI auditing literature, conceptualizations of accountability are linked to fairness and transparency. Nevertheless, fairness, transparency, and accountability often are viewed as separate principles [18,42,47,62,63,82,100,102,123,133,138,147]. However, within the broader concept of accountability, transparency is recognized as one of the five central principles alongside responsiveness, responsibility, remediability, and verifiability [121]. Furthermore, accountability is acknowledged as one benefit of transparency [70,98]. It is also closely associated with responsibility, liability, and transparency, as demonstrated by its connection to the European Union's General Data Protection Regulation (GDPR) [135]. Accountability has been widely considered in the GDPR, as the regulation's articles imply that data controllers are responsible for leading the compliance effort [145].

Properly designed algorithmic audits are vital for better accountability [46]. The lack of a legal accountability mechanism is argued to be one of the main weaknesses of a principled approach to AI ethics [96].

As governments adopt algorithms to support decision-making processes, it has become more urgent to address issues of responsible use and accountability [129]. Recently, the European Commission has proposed the Artificial Intelligence Act as a framework for preventing, reporting on, and allocating accountability for different kinds of system failures [102,138].

To complement auditing, explainable AI tools and features have been identified as one solution to achieve more accountable AI because they make AI systems' decisions and underlying reasons more transparent for users [98]. An explainable AI system can be defined as a self-explanatory, intelligent system that describes the reasoning behind its decisions and predictions [98]. Other solutions include employing auditing practices and data audits that enforce data lineage and accountability to help organizations meet strict regulatory requirements and benefit from an overarching perspective on their data assets [147].

Conceptually, accountability-improving methods can be separated into three stages: ex-ante, in-ante, and post-ante [63]. Ex-ante methods (e.g., impact assessments) focus on algorithm development and deal mainly with the algorithms' planning and design phase, in-ante methods implement accountability measures in the development lifecycle, and post-ante methods provide accountability measures after the model is deployed. [63].

4.2.5. *Privacy (25 studies)*

Privacy issues include challenges and values associated with data protection and securing private and personal information [59]. In general terms, privacy "makes sure that the sensitive data that is either shared by an individual or collected by an AI system is protected from any unjustifiable or illegal gathering and use of data" [63]. From an ethical perspective, privacy involves risks related to sensitive data such as juvenile and biometric face data [118]. Privacy is also a critical factor in LaBrie and Steinke's [76] proposed ethical AI algorithm audit framework, which combined the privacy, accuracy, property, and accessibility framework with a layered AI governance model [43].

AI system operations and AI auditing both require data, and large data sets can present privacy risks for individuals represented in the data set. Private organizations, governments, or hackers could misuse personal data, leading to harmful consequences [63]. Data privacy protects individuals from being identified or associated with certain information because automated decision-making systems may leave decision subjects vulnerable to invasions of privacy [102,147]. However, auditing data sets may include data on those impacted by the audited technology. Therefore, the privacy principle can set a contradictory challenge for ethics-based AI auditing. Sensitive and biometric information may be stored, and there are risks that these data sets can be accessible beyond the intended auditing purpose; hence, it is crucial for AI systems to protect users' privacy [63,69,118].

Privacy is conceptually linked with fairness and discrimination, as fairness can be viewed as hiding information. Fairness keeps specific attributes, e.g., race and gender, private when a fair decision entails not allowing inferences based on a decision subject's attributes. Thus, the main privacy challenges of fairness auditing come from restrictions on collecting sensitive variables [25,69,114,115].

ML is based on making complex models using data, and different techniques are used to manage privacy in data analytics. For example, Domingo-Ferrer et al. [32] presented a methodology to let individual subjects on whom automated decisions have been made elicit a rule-based approximation of the model underlying the decision algorithm. The methodology seeks to offer a solution to the so-called privacy-accuracy tradeoff. Privacy and data protection are also key components of various legal frameworks and part of the legal requirements for accountability [135]. However, privacy harms have been accused of being murky and vague, obscuring privacy boundaries and hampering attempts to contextualize discussions within the general legal theory of privacy [83].

4.2.6. *Trust (22 studies)*

Trust is presumed necessary for AI adoption [66], but various scandals over biased outcomes, transparency issues, and data misuse have led to a growing mistrust of AI [16]. Reliance on search engines and its effects on democracy are one striking issue related to trust [122] because many people use Google as their primary fact-checking tool. The question of trust has amplified calls for ethical audits of algorithms and embedding ethical principles in AI practices to increase public trust in technology [3,40]. In the literature, trust was conceptualized chiefly via trustworthy AI, most prominently in the AI HLEG's "Ethics Guidelines for Trustworthy AI," which were used in this paper to differentiate ethics-based from non-ethics-based articles [1,16,63,100,102]. Mohseni et al. [98] defined trustworthiness as "enabling positive user attitudes toward the system that emerges from knowledge, experience, and emotion," while Kaur et al. [63] defined trustworthy AI as a framework to ensure that a system is worthy of being trusted based on the evidence concerning its stated requirements. It ensures that users' and stakeholders' expectations are met in a verifiable way. Different disciplines define trust in many ways, but various definitions agree that trust involves integrity and reliability [63].

Extant research has found that people generally perceive AI evaluation as less trustworthy than human evaluation [114]. Explicability is highlighted as necessary for ensuring public trust and technology understandability. The vision is to develop AI technology in a way that secures people's trust while serving the public interest and strengthening social responsibility [3,40]. Attaining this vision requires that society develop an accessible redress mechanism for harms inflicted, costs incurred, and other technology-driven grievances.

Many ethics-based AI auditing principles are related to trust, but the principles often leave a gap between the "what" and "how" of AI development [3]. For example, fairness, explainability, accountability, privacy, and acceptance are seen in the literature as critical requirements of trustworthy AI [63,109]. Regulations and transparency also could achieve greater trust in AI systems [37,63,77,149]. However, more transparency does not always entail more trustworthiness, and systems should be designed for optimal transparency [115]. Shneiderman [133] proposed recommendations on three governance levels—team, organization, and industry—to increase reliability, safety, and trustworthiness. Governance structures on these levels clarify who takes action and who is responsible, connecting to the responsibility principle discussed previously.

New auditing tools and frameworks aim to increase trust in AI systems. For example, the INFER framework seeks to improve recommender systems' trustworthiness by combining explainability, fairness, and user interaction research [45], and Park et al. [115] created a fairness audit framework that assesses ML algorithms' fairness, focusing on security issues, e.g., trustworthiness.

4.2.7. *Beneficence (20 studies)*

Beneficence (the promotion of well-being and other beneficial outcomes) is rarely explicitly defined. Beneficence can be viewed as near-equivalent to non-maleficence because "do only good" and "do no harm" represent similar values. In the AI auditing literature, beneficence is mainly linked with other ethical principles to highlight the benefits that the other principles can elicit, e.g., privacy and accountability [123], transparency [82], or other technical and legal benefits for system designers, operators, auditors, regulators, and end-users that come from improved accountability [132,135].

Although beneficence is a broad and somewhat abstract principle, it typically is featured at the top of lists of principles [93,102]. Examples include the Montreal Declaration for Responsible AI, which states that "the development of AI should ultimately promote the well-being of all sentient creatures," and the IEEE's Ethically Aligned Design principles, which proclaim that "we need to prioritize human well-being as an outcome in all system designs" [40]. Many fields—e.g., healthcare, education, and security—have high hopes for the benefits of AI alongside discussions of its potentially harmful consequences. Thus, beneficence can be characterized as an abstract and general goal of auditing rather than a principle to be audited directly.

According to the literature, users of ML systems are likely to seek out technology with a suitable accountability mechanism and high social benefits [82,121]. Reliable, safe, and trustworthy technologies are seen to benefit individuals, organizations, and society [63,133]. Auditing algorithms can benefit both data controllers and data subjects [83], and AI systems can support socially and environmentally beneficial outcomes [101].

4.2.8. Freedom/autonomy (13 studies)

Explicit conceptualizations of freedom and autonomy in the context of AI auditing are rare. Reed et al. [[121], p. 17] defined autonomy as a "fundamental principle under which individuals should be entitled to make their own choices as an act of will, rather than having those choices made for them and forced upon them." In bioethics, the idea of freedom and autonomy is that individuals should have a right to make their own decisions about the treatment they do or do not receive [40]. However, autonomy in ML systems is problematic because their choices are based on what they have learned rather than principles, and choices often are invisible to technology users [121]. In this sense, ML systems differ from rule-based IT systems and the medical treatments that bioethicists have discussed (cf. [40,96]).

In the AI auditing literature, autonomy is an underlying principle rather than an audited criterion. The AI4People initiative analyzed guideline documents, four of which mentioned the principle of autonomy, promoting the importance of humans' autonomy to set their own norms [40]. The "everyday algorithm auditing" framework [132] places users' autonomy and agency at the center because they play a major role in deciding their own course of action. The risks and impacts of possible incorrect decisions mean humans should be involved in decision-making [63].

4.3. Stakeholders of AI auditing

The second research question focuses on the stakeholders of AI auditing by asking *what knowledge contributions the AI auditing literature presents to stakeholders*. The main stakeholders to whom studies target AI auditing tools or concepts are system developers and deployers, the wider public, researchers, auditors, AI system users, and regulators. Different stakeholders include AI auditing in their work; therefore, many actors are connected to auditing processes. Indirectly, AI auditing may affect all citizens. However, for clarity, we focus on the stakeholders presented in the reviewed studies. We identified the stakeholders primarily by searching for explicit mentions of target groups in the studies. If these were not found, we inferred the target group from the paper's content more broadly. Table 5 lists the stakeholder groups targeted in the reviewed studies, and the following sections elaborate on how each stakeholder group is discussed. We include auditors as stakeholders because they are mentioned as target groups in the AI auditing studies. Of course, auditors are also the actors carrying out auditing and, thus, considering all the stakeholders.

Table 5
AI auditing stakeholders targeted in the reviewed studies.

Stakeholders	Studies
System developers and deployers	P1, P2, P4, P14, P16, P30, P33, P38, P39, P41, P43, P44, P45, P47, P48, P54, P55, P57, P60, P62, P70, P72, P73, P74, P75, P77, P80, P83, P84, P85, P90
The wider public	P3, P6, P12, P14, P17, P25, P28, P29, P35, P38, P52, P54, P55, P56, P60, P62, P64, P68, P70, P71, P73, P79, P84, P86, P87, P88, P90, P91, P92, P93
Researchers	P4, P9, P21, P22, P25, P31 P33, P39, P40, P42, P45, P47, P48, P49, P50, P51, P52, P55, P61, P73, P75, P79, P81, P92, P93
Auditors	P2, P3, P7, P10, P13, P14, P15, P18, P23, P24, P26, P28, P37, P57, P58, P59, P64, P65, P68, P69, P71, P78, P89
AI system users	P8, P11, P20, P27, P36, P39, P48, P51, P60, P66, P67, P76, P79, P83, P85, P89, P93
Regulators	P10, P14, P34, P35, P36, P43, P44, P45, P56, P65, P69, P71, P72, P74, P88, P90

4.3.1. System developers and deployers (31 studies)

Studies focusing on developers and deployers provide tools and procedures for developers to audit the algorithms that support AI systems [4,6,10,20,25,37,42,66,73,76,77,109,118,138,149]. They also help ML development and deployment processes recognize and avoid potential liability issues and other harmful behaviors [95,100,120,132,136,147], as well as providing recommendations and ethics codes for developing AI [3,27,40,47,88,112]. The main objective of studies targeting system development and deployment actors is to provide them with the best tools and procedures for conducting ethics-based AI auditing, as well as understanding how algorithms treat people and how to audit them.

Collaboration between developers and algorithmic auditors can lead to more ethically acceptable technologies. AI developers are often not competent or trained enough to address algorithmic fairness, accountability, and transparency issues, or they do not know how to use correct methods to identify potential discrimination [4]. Regarding tools and procedures for AI auditing, the SMACTR framework ensures audit integrity in system development and deployment and pre-empts negative consequences by presenting mechanisms that help developers meet ethical expectations and standards in AI systems [118]. Furthermore, available tools include the visual analytics system Deblinder [20]; the FairLens tool for discovering and explaining biases and auditing black-box models [109]; and a matrix for auditing algorithmic decision-making systems [138]. Another toolkit aimed at developers is AI Fairness 360 by Bellamy et al. [10], which enables developers to improve new algorithms and use the toolkit for benchmarking. Concerning guidance and the provision of recommendations and ethics guidelines, Floridi et al. [40] listed ethical principles that developers should adopt in their AI development process to contribute to a better AI society. Principles and recommendations adopted in the AI development process serve all the stakeholders and increase public trust and acceptance of the development process. For the same purpose, Grasso et al. [47] presented their framework for developers, demonstrating how compounding accountability frameworks and domain-specific codes of ethics can help answer ethical expectations in AI systems.

4.3.2. The wider public (30 studies)

The motivation for ethics-based AI auditing is mostly on individuals, groups, or companies indirectly benefiting from AI auditing. Even when the auditing tools, practices, or frameworks were targeting, e.g., developers or auditors, the main objective was to ensure fair, unbiased, and transparent treatment of people in general [18,21,24,37,63,70,111,112,120,122,123,129,133] and create systems that can benefit individuals and companies [55,109,115,149]. Auditing tools also strive to ensure that individuals understand decision-making systems' functionality, impacts, and consequences [42,45,46,83,116]. Thus, the overall intention is to design human-centered, reliable, safe, and trustworthy AI that benefits individuals and companies [66,77,100,132,133].

Legislation is a central driver in ensuring respect for individual rights, and the GDPR notably offers an initial regulatory framework for automated individual decision-making [42]. Its provisions supply information on the impacts and consequences of decision-making systems for individuals [83]. Legibility-by-design systems promoted by the GDPR inform individuals about the existence and logic of the system's functionality and decision-making algorithms' specific decisions [83]. The GDPR aims to tackle the unfavorable treatment that individuals or groups receive based on any special categories [46].

As an example of ensuring unbiased treatment, biases in word embeddings lead to algorithmic discrimination toward social groups and individuals [111], and it has been demonstrated that discrimination influences hiring processes [21]. AI auditing helps understand issues and identify needed measures to ensure fair algorithms [111], and it plays a vital role in ensuring unbiased treatment [21].

4.3.3. Researchers (25 studies)

Researchers lay the groundwork for developers to conduct AI auditing and understand machine behavior [4,6,25,73,98,129,146]. Moreover, they guide empirical research on how to study algorithmic discrimination and improve algorithmic fairness, as well as other unique ethical considerations [37,58,77,124,134,137,141,149,63,70]. They also create methods to improve AI auditing [10,21,26,37,49,109,141]. In addition, researchers have demonstrated the difficulties in studying fairness in ML systems, e.g., the tensions affecting social science audits [146], and proposed recommendations that have laid the groundwork for further investigation.

Among this groundwork, insights can be found on awareness of harms and biases in terms of dealing with issues of fairness in image analysis algorithms [4,73], as well as systematically analyzing the ML pipeline, with an emphasis on visual privacy and bias issues [25]. To guide empirical researchers on ethical consideration issues, researchers have developed frameworks to understand AI fairness and bias by presenting crucial audit components across categories [77] and to build and audit fair algorithms and structure audits to be practical, independent, and constructive [149]. Considering methods to improve AI auditing, researchers have adapted existing bias-detection mechanisms to provide access to auditing ML systems' decision processes and improve existing systems [141] and developed a publicly available methodology for understanding machine behaviors and AI auditing [6].

4.3.4. Auditors (23 studies)

AI auditors, i.e., auditing professionals, are a key stakeholder group for AI auditing studies in addition to being the main actors conducting the auditing. Studies aimed at auditors create frameworks or system architectures for better auditing practices [36,51,55,65,76,84,145], as well as examine social networks' auditing elements [82,122,125,144]. They also identify ethical problems related to facial recognition systems [4,34,118,151], as well as guide auditing to minimize harms and improve trustworthiness [2,3,29,100]. In terms of frameworks for auditing practices, some frameworks aimed, for example, to reduce discrimination in job markets and ads, as well as automate parts of AI auditing in HR [55,84], help auditors identify why mistakes happen in the predictor model, produce examples of inputs in which the predictor is erring significantly [65], develop new norms, and help auditors clarify the audit scope limit, as auditors need to live up to ethical ideals [118].

4.3.5. AI system users (17 studies)

AI system users were another identified stakeholder group in ethics-based AI auditing. Studies focusing on users aimed to present tools that allow users to audit ML models by themselves [10,19,62,150]. Moreover, they presented frameworks that enable users to be more aware of potential biases and give users more control [8,32,45,71,72,132]. These studies also developed auditing instruments to serve the interests of all users and stakeholders that the algorithms affected [16,20,29,88], as well as aimed to increase non-expert users' understanding of systems [63,134].

Regarding tools for users to audit ML models, AI Fairness 360, FAIRVIS, and Algorithmic Equity Toolkit, among others, help users understand how ML works, determine whether the system is driven by AI, and ask questions about the context, allowing users to reach their assessments and go from raw data to a fair model and, thus, elicit results [10,19,62]. For industrial or intellectual property reasons and due to different user competence levels, open-source algorithms cannot always be the solution, and users cannot always have full knowledge about algorithms. Therefore, Domingo-Ferrer et al. [32] presented an approach in which users affected by an AI system, in a collaborative way, can make a rule-based approximation of the model underlying the decision algorithm. Similarly, Kulshrestha et al. [72] targeted search engines to increase transparency and decrease discrimination against users, and Shen et al. [132] proposed the concept of everyday algorithm auditing in which users detect, understand, and interrogate machine behaviors in their interactions with the system.

4.3.6. Regulators (16 studies)

Appropriate regulation allows AI to reach its potential, reducing fear, ignorance, and misplaced concerns about AI, and widely accessible regulatory mechanisms increase public acceptance and adoption of AI technologies. Studies that focused on regulators have considered the EU GDPR [42,46,123,145], legal algorithm standards, and processes for regulatory oversight [16,24,51,138], as well as examined whether current regulations correspond with the ethical principles and current state of AI systems [40,66,96,100,102]. According to Mittelstadt [96], a unified regulatory framework does not exist in the AI field, although the recently proposed European Union AI Act is an attempt at consolidate regulation [102]. Therefore, many reviewed studies aimed to advance legal mechanisms for regulators and regulatory frameworks for AI development. For example, Cobbe et al. [24] provided a reviewability framework that draws on administrative law's approach to reviewing human decision-making and offered a practical way forward toward a more legally relevant form of accountability for automated decision-making.

Floridi et al. [40] encouraged the inclusion of ethical, legal, and social considerations in AI projects to answer both ethics and policy calls. They also highlighted self-regulatory codes of conduct, as many current attention manipulation techniques can be constrained through them [40]. AI companies and developers are committing themselves to ethical principles and self-regulation codes, which might lead to policymakers not pursuing new regulations [96]. For example, processing personal data requires following strict criteria, which is why the EU GDPR prompts companies and organizations to audit their services [42]. The GDPR focuses on protecting personal data, but it also provides guidance to address algorithmic decision-making's effects, paving the way for third-party inspections of AI auditing. However, the GDPR does not specify who should conduct audits [46].

Ethics-based auditing should inform, formalize, assess, and interlink existing governance structures [100]. Regulators could promote ethics-based AI auditing by providing standardized reporting formats, facilitating knowledge exchange, providing guidance on normative tensions, and creating an independent body to oversee ethics-based auditing of automated decision-making systems [100]. Several recent regulatory proposals have called for algorithm audits [138], particularly the proposed European Artificial Intelligence Act, the first general legal framework for AI [102].

5. Discussion and conclusion

This study aimed to understand and consolidate the current literature on ethics-based AI auditing by synthesizing conceptualizations of ethical principles and outlining knowledge contributions to stakeholder groups. We addressed two research questions: 1) How does the academic ethics-based AI auditing literature conceptualize AI ethics principles, and 2) What knowledge contributions does the literature present to the stakeholders of ethics-based AI auditing? To answer the first research question, from the sample of 93 ethics-based AI auditing articles, we synthesized the most important ethical principles in ethics-based AI auditing literature. Fairness, transparency, non-maleficence, and responsibility were the most common ethical principles, followed by privacy, trust, beneficence, and freedom/autonomy.

Addressing the second research question, we identified the most critical stakeholders in AI auditing: AI system developers and deployers; the wider public; researchers; auditors; AI system users; and regulators. We determined the different knowledge contributions that ethics-based auditing studies offer each stakeholder group in terms of guidance, tools, and frameworks.

Building on these findings, this study provides two synthesizing insights to the literature on ethics-based AI auditing (e.g., [16,102,128]): 1) a refined understanding of ethical principles in the AI auditing literature by demonstrating their prevalence and the conceptualizations of focal concepts and 2) clarification of the knowledge contributions for different stakeholders in ethics-based AI auditing studies. These insights seek to consolidate the disconnected literature streams in ethics-based AI auditing, and they are discussed in the following sections before turning to the implications for the relevant IS literature streams.

5.1. Conceptualizations of focal concepts

This study's first synthesizing insight is a detailed review of the current literature on ethics-based AI auditing, providing an empirical grounding to discern what the literature currently discusses. This includes descriptive data on research methods and main outputs. The AI auditing literature includes empirical (including design science) and conceptual studies (see Appendix B), with empirical studies being the most prominent. In terms of the main principles in the ethics-based AI auditing literature, the findings demonstrated that the clear majority of the studies considered fairness-related issues, followed by transparency, responsibility, and non-maleficence.

The results revealed an emerging convergence around certain ethical principles, e.g., fairness and transparency. The review also found that studies addressing non-maleficence significantly outnumber those addressing beneficence, implying that ethics-based AI auditing is more concerned about preventing harms than highlighting benefits. Notably, the reviewed studies did not address sustainability, dignity, and solidarity concepts. Ethical concerns related to individuals and delimited social groups (e.g., fairness) were brought to the forefront, which might explain why, for example, environmental and solidarity aspects were not considered.

Conceptualizations of the key concepts of fairness, bias, transparency, and accountability vary significantly within the reviewed literature. Most of the principles have no established definitions, and they are conceptualized in different ways in the studies. This review contributes to the discussion by bringing together different conceptualizations and highlighting the concepts used in the literature (see Tables 6 and 7). In Table 6, similar conceptualizations are grouped for the sake of conciseness. This summary indicates that a significant demand exists for the systematization of definitions and, eventually, metrics in the field of ethics-based AI auditing. In the subsequent Table 7, we systematically categorized and analyzed various AI ethics definitions, offering a structured perspective on technical and social aspects, and distinguishing between process-driven and outcome-driven approaches, thus providing an understanding of the multifaceted ethical dimensions of AI.

The literature on fairness highlights its multiple definitions. The complexity of the fairness concept necessitates different fairness metrics. Metrics also might vary depending on culture or context. For example, fairness metrics might include measuring a model's group-specific false-positive rates, calibration, and other metrics, and researchers or users need to decide which metrics to examine. The most common distinction is between group fairness and individual or procedural fairness [6,10,25,50,102,125]. Group fairness examines differences across groups and whether groups have parity in terms of benefits received. However, individual fairness concerns the process by which individuals are judged and that similar individuals receive similar treatments. Procedural fairness is defined as involving interpretability methods that attempt to understand how a prediction is made [119]. Other fairness conceptualizations include codes of ethics, algorithm-driven actions, and processing data that might lead to negative consequences for vulnerable populations [26,71,85]. Therefore, fairness can be viewed as a lack of any prejudice or favoritism toward an individual or group based on

Table 6
Conceptualizations of ethical principles in AI auditing.

Terms and authors	Synthesis of conceptualizations
Fairness	
(F1) [49,50]	Fairness definitions address a narrow set of considerations, as they mostly are developed to address specific problems, and different definitions include different metrics. The concept of fairness can be understood in reference to the different social groups that comprise the organization of society.
(F2) [6,10,50,85,102,125,25]	Fairness definitions can be separated into group fairness (examining differences across groups and that groups have parity in terms of benefits received) and procedural or individual fairness (concerning the process by which individuals are judged and similar individuals receive similar treatment).
(F3) [6,10,19,47,109,111,150]	Codes of ethics offer tools for defining fairness metrics, even though fairness metrics vary between applications, even in a specific context. Different fairness metrics depend on context and culture, and fairness definitions include different metrics.
(F4) [64,85,116]	Most common fairness definitions are statistical, which proceed by fixing a small number of "protected subgroups," then asking that some statistic of interest be equalized approximately across groups. Popular measures include statistical parity, conditional statistical parity, false positive and false negative rates, and predictive parity.
(F5) [63,77,95,141]	Fairness entails treating subjects similarly regardless of their defined protected attributes. No unique definitions exist; therefore, they are based on individual perspectives and definitions of fairness. Actions driven by algorithms can be assessed according to numerous ethical criteria and principles, which we generally refer to as "fair actions." Fairness of the system ensures the absence of any discrimination or favoritism toward an individual or group.
(F6) [119]	"Procedural fairness" for machine learning (ML) systems involves interpretability methods that attempt to understand how a prediction is made.
(F7) [26,71,85]	Fairness is motivated by the concern that high-impact decisions that machine-learned systems make may have negative consequences for vulnerable populations. Fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.
(F8) [137]	Fairness can be classified broadly into varieties that involve pre-processing data that go into the algorithms, processing during the prediction algorithms themselves, and those that post-process the results of an existing algorithm to allow for fairer decisions.
(F9) [4,73]	Fairness "is best understood as a placeholder term for a variety of normative egalitarian considerations."
(F10) [134]	Three major categories of fairness conceptualizations are: (1) It may be understood through the lens of individual attitudes; (2) it may be understood through the lens of legality, ethicality, and morality; (3) it may be understood from various domain-embedded technical perspectives in which different research communities have their own technical definitions of each term.
Bias	
(B11) [83,126,141]	Machine biases refer to discriminatory practices or consumers' misrepresentations that can be classified into "cognitive biases" and "statistical biases." Cognitive biases occur when data collection errors lead to inaccurate depictions of reality, and statistical biases occur when the underlying subject matter draws on information that also is linked inextricably to structural discrimination. Biases originate either from data used in training a machine learning algorithm or the algorithm itself, which should be controllable as bias in the databases used to train and evaluate machine learning algorithms.
(B12) [19]	Naturally encoded societal biases in the ML models often are referred to as algorithmic biases.
(B13) [10,25,42,71,85,111]	Algorithmic discrimination or algorithmic bias is defined as disadvantageous differential treatment of (or impact on) an already disadvantaged group. The criteria by which what constitutes bias is defined also need to be framed from multiple perspectives, with as many as 23 different types of

(*continued on next page*)

Table 6 (*continued*)

Terms and authors	Synthesis of conceptualizations
	bias definitions in existence, and data bias can exist in many shapes and forms, with different bias-handling algorithms addressing different parts of the model lifecycle.
(B14) [111]	Algorithmic bias denotes deviation of the algorithmic results from specific social expectations based on epistemic or normative reasons. Bias might result in unfair or discriminatory decisions and statements in three types 1. preexisting (input data); 2. technical (software, hardware, or mathematical constraints); and 3. emergent (evaluation of results and the context of their application).
(B15) [73]	A "biased algorithm" yields outputs that deviate systematically from what is expected.
(B16) [63]	Three types of bias are data bias (if the data on which the system is trained is biased), model bias (if the algorithm itself introduces it), and evaluation bias (if the wrong evaluation metrics were used to evaluate the model).
(B17) [88]	Bias is defined broadly as a skew that produces harm, including the subcategory of representational harm: when a system perpetuates or exaggerates social inequalities along identity lines.
(B18) [37]	A prejudice or tendency in predictions made by an AI-DMS leading to decisions against or in favor of one individual or group in a way considered to be unfair.
Transparency	
(T19) [18,95,136]	Transparency provides information on the demographic and phenotypic composition of training and benchmark data sets, referring to the degree to which algorithms can be seen and understood, information availability, and how information may support the user's decision-making process.
(T20) [121]	The property of a system, organization, or individual of providing visibility of its governing norms, behavior, and compliance of behavior to the norms—an aspect of the wider concept of accountability.
(T21) [46,82]	Algorithm transparency introduces the "right to explanation," whereby data subjects are entitled to "meaningful information about the logic involved, as well as the significance and the envisaged consequences" when automated decision-making or profiling takes place. Control transparency is a way to make information accessible and to communicate for any purpose.
(T22) [123]	Transparency is understood as the availability of information without loss, noise, delay, or distortion.
(T23) [63]	An AI system's transparency refers to the need to explain, interpret, and reproduce its decisions. It ensures that the various stakeholders using or impacted by the system clearly understand its performance and limitations.
(T24) [134]	The degree to which the model is understandable by itself; mainly refers to the model's characteristics.
Accountability	
(A25) [118,145]	The state of being responsible or answerable for a system, its behavior, and its potential impacts.
(A26) [135,147]	"Algorithmic accountability" concerns issues of fairness, transparency, and explainability, particularly regarding machine learning. Accountability involves apportioning responsibility for a particular occurrence and determining who owes any explanation for that occurrence.
(A27) [18]	We define accountability as reporting algorithmic performance on demographic and phenotypic subgroups and actively working to close performance gaps where they arise.
(A28) [121]	Five core accountability attributes are transparency, responsiveness, responsibility, remediability, and verifiability.
(A29) [47,63]	Algorithmic accountability includes an action plan for redress when things go wrong, as well as incentives for the iterative development of algorithmic systems with the inevitable evolution of their intended domain. AI systems should be able to justify their decisions.
(A30) [82,123]	Any definition of accountability will include at least three elements: (1) responsibility; (2) answerability; and (3) sanctionability of the accountable party.
(A31) [123]	Information accountability "means that the use of information must be transparent in order to be able to determine whether a given use is appropriate under a set of rules, and that the system allows individuals and institutions to be held responsible [to be held accountable] for misuse."

Table 7
Categorization of ethical principles' conceptualizations in AI auditing.

	Social aspects	Technical aspects
Process-oriented approach	AI auditing as interrogating how complex social groups and contexts are represented in algorithmic systems Conceptualizations: F1, F2, F3, F10, B12, T19, T20, T21, T23, A26, A27, A28, A29, A30	AI auditing as technically governing algorithmic processing Conceptualizations: F6, F8, B11, B14, B16, T22, T24, A31
Outcome-oriented approach	AI auditing as avoiding social discrimination, prejudice, and harm in predictions and decisions Conceptualizations: F5, F7, F9, B13, B17, B18, A25	AI auditing as technically assuring appropriate outputs Conceptualizations: F4, B15

inherent or acquired characteristics [26,71,85].

Like fairness, bias is conceptualized in various ways. The most common classification is between cognitive bias and statistical bias [83,126,141]. Cognitive bias describes errors in data collection that lead to inaccurate depictions of reality, while statistical bias happens when the underlying subject matter draws on information that also is linked inextricably to structural discrimination. More generally, bias is defined as disadvantageous differential treatment of (or impact on) an already-disadvantaged group [10,25,42,71,111]. However, Mehrabi et al. [85] note that bias in data can exist in many forms and listed 23 types of bias, including preexisting, technical, and emergent dimensions.

Transparency, in turn, is most commonly defined as the degree of availability of information in systems that helps users understand them and support their decision-making processes [18,95,136]. Transparency pertains to meaningful information about the logic and consequences of systems when decision-making is automated. In addition, it is seen to fall under the broader concept of accountability, providing visibility to the systems governing norms, behavior, and compliance of behavior to norms, as well as fulfilling a regulatory requirement to assign responsibility or liability, thereby facilitating accountability and providing explanations for users [121].

Accountability is also conceptualized in several distinct ways. The most general definition of accountability refers to being responsible or answerable for a system, its behavior, and its potential impacts [118,145]. However, accountability is also viewed as reporting algorithmic performance on demographic and phenotypic subgroups and actively working to close performance gaps wherever they arise [18]. Accountability is linked closely to other AI auditing principles. For example, Singh et al. [135] state that algorithmic accountability concerns fairness, transparency, and explainability. It involves assigning responsibility for events and determining who is owed any explanations for such occurrences.

In Table 6, we presented various conceptualizations of ethical principles in AI auditing. Building on that foundation, Table 7 offers a systematic categorization of these distinct interpretations, together with a synthesis of how AI auditing is viewed within specific categories of conceptualizations. Through this organization, we aim to clarify the commonalities and differences among the conceptualizations and, subsequently, tease out different underlying understandings of AI auditing. This matrix offers a lens to view how the literature approaches AI auditing as a socio-technical and multi-dimensional phenomenon. By visually mapping these conceptualizations, we can begin to consolidate the fragmented landscape of AI auditing and identify areas of overlap, distinction, and potential gaps.

The matrix utilized for analyzing various conceptualizations related to AI ethics categorizes them based on two primary dimensions: "social aspects" and "technical aspects," as well as "process-oriented" and "outcome-oriented" approaches. Each conceptualization is placed based on its primary emphasis and orientation, and both dimensions are explained further later. Many conceptualizations could fit into multiple categories, but for the sake of clarity, they are placed where they align most strongly.

The "Technical aspects/Social aspects" dimension categorizes definitions based on whether they focus more on the technical, mathematical, or algorithmic components (technical aspects) or whether they address broader social, cultural, ethical, or normative considerations (social aspects). Social aspects move beyond the algorithm to address broader social, ethical, and cultural issues as well as human-centric perspectives. They encompass societal impacts, normative values, and ethical considerations. For instance, a definition that discusses bias in terms of societal or cultural discrimination falls into this category. Technical aspects predominantly speak around the machinery, algorithms, data, and specific technical nuances of AI systems or the algorithmic, mathematical, or structured components of the principle in question. They might reference specific methods, models, or statistical measures. For example, a definition of fairness in terms of statistical parity would be categorized here.

The "Process-oriented/Outcome-oriented" dimension categorizes conceptualizations based on whether they address the methods, procedures, or operations of AI (process-oriented) or are concerned with the results, effects, or consequences of AI operations (outcome-oriented). Process-oriented conceptualizations emphasize the mechanisms, methodologies, and steps taken within AI practices. This category centers on the methodologies, mechanisms, techniques, and actions to achieve a specific goal or objective. It focuses on the "how" aspect. These conceptualizations might discuss how an algorithm is designed, how data is processed, or how decisions are made. For instance, in AI fairness, a process-oriented approach might involve methods or steps to ensure that an AI system is trained fairly. The outcome-oriented category pertains to the results, consequences, or effects of a particular action or series of actions. Outcome-oriented conceptualizations might discuss the effects on individuals or groups, the broader societal implications, or the tangible outcomes of algorithmic decisions. It emphasizes the end result or the "what" aspect. In the context of AI fairness, an outcome-oriented approach might describe the desired equitable result of an AI system's decision-making or the consequences of its actions.

The social aspects and process-oriented (S&P) cell appeared as the most dominant cell. Conceptualizations here consider AI auditing as the interrogation of how complex social groups and contexts are represented in algorithmic systems. The emphasis on this cell implies that a significant portion of AI ethical considerations is about understanding how systems are designed, trained, and implemented, ensuring they adhere to socially acceptable norms and practices, even before tangible results manifest. These definitions often address the emergence, sources, and means of addressing biases and unfairness, especially as they relate to societal groupings and demographic differences, without necessarily pinpointing specific harmful outcomes. The focus is on integrating societal norms, values, and fairness at every stage of AI development and implementation.

With fewer conceptualizations than the S&P category but still a significant number, the social aspects and outcome-oriented cell underscores the importance of directly measurable impacts of AI on society. From this perspective, AI auditing is viewed as avoiding social discrimination, prejudice, and harm in predictions and decisions. The emphasis here reflects a concern for the tangible real-world consequences of algorithmic processes and the need to ensure they align with societal values and expectations. There seems to be an acknowledgment that despite best efforts in the process, undesired outcomes can still emerge. As a result, there's a recognized need to continually measure and correct AI's societal impact post-deployment, ensuring dynamic adaptability to evolving societal standards.

The high presence of definitions in the technical and process-oriented cell speaks to the foundational importance of the mechanics behind AI systems, unbiased data, and the methodologies employed in AI. Here, AI auditing means technically governing algorithmic processing. Although societal implications are crucial, this category's emphasis reminds us that understanding and refining the technical processes is foundational to achieving desired social and technical outcomes. The relatively smaller number of definitions here, compared to S&P, may indicate more agreement on what constitutes good technical practices or, alternatively, their overshadowing by more pressing social concerns in the current ethics-based AI auditing discourse.

The technical aspects and outcome-oriented quadrant had the fewest conceptualizations. From this perspective, AI auditing is seen as technically assuring appropriate outputs. The few conceptualizations suggest technical and outcome-oriented approaches might be more straightforward or less debated in current ethics-based AI auditing discussions. The focus might be more on ensuring technical processes are robust rather than just the outcomes, implying that a well-implemented process might inherently lead to a good technical outcome. Alternatively, the lack of definitions might indicate a need for more discourse on direct technical outcomes in relation to AI ethics.

Overall, process-oriented approaches dominated the matrix. This might be because fairness, transparency, and accountability involve processes and actions that need to be consistently applied and evaluated rather than merely outcomes to be achieved once. Furthermore, the complex nature of AI algorithms means that once an undesirable outcome is observed, it can be challenging to trace back to the root cause or adjust the system without a clear process in place. Outcomes can also sometimes be ambiguous and debated. Process-oriented approaches focus on ensuring that AI systems are developed and deployed ethically from the start, which can help anticipate and prevent issues before they arise. A strong process orientation also ensures that as the models adapt and learn, they continue to adhere to ethical guidelines.

5.2. Stakeholders and knowledge contributions of ethics-based ai auditing

This study's second synthesizing insight is a framework summarizing the stakeholder groups of AI auditing and what the studies offer to each of these. AI auditing studies' targeted stakeholder groups and the knowledge contributions for these groups (Sections 4.3.1–4.3.6) are

Table 8
Stakeholder groups and the knowledge contributions of ethics-based AI auditing.

Stakeholder group	Knowledge contributions
System developers and deployers	- Tools and procedures [4,6,10,20,25,37,42,66,73,76,77,109,118,138,149] - Guidance for development and deployment to recognize and avoid harms [95,100,120,132,136,147] - Recommendations and ethics codes for developing AI [3,27,40,47,88,112]
The wider public	- Ensuring fair, unbiased, and transparent treatment [18,21,24,37,63,70,111,112,120,122,123,129,133] - Ensuring individuals' understanding of AI systems [42,45,46,83,116]
Researchers	- Groundwork for AI auditing [4,6,25,73,98,129,146] - Methods and methodologies [10,21,26,37,49,109,141] - Guidance on conceptualizing issues [37,58,63,70,77,124,134,137,141,149]
Auditors	- Frameworks for auditing practices [36,51,55,65,76,84,145] - Studying auditing elements of algorithmic systems [5,34,82,118,122,125,144,151] - Guide auditing to minimize harms and improve trustworthiness [2,3,29,100]
AI system users	- Tools to audit ML systems [10,19,62,150] - Enabling awareness of biases and giving control [8,32,45,71,72,132]
Regulators	- Consideration of regulation and legal standards [16,24,42,46,51,123,138,145] - Correspondence between regulation, principles, and AI systems [40,66,96,100,102]

summarized in Table 8, and we synthesize these contributions before offering some critical remarks on gaps in the literature.

Considering what AI auditing provides to stakeholders, we can synthesize the knowledge contributions into three types: cognitive, pragmatic, and normative. These types of knowledge contributions depict the orientations of the studies toward stakeholders, i.e., whether they provide guiding knowledge (cognitive), operationalization in terms of tools and methods (pragmatic), or empowerment that enhances the capabilities of stakeholders as ethical agents (normative). First, with respect to the cognitive knowledge contributions, *guidance* provided to regulators, researchers, and developers is the most information-intensive of the knowledge contribution types, for example, assisting them in making sense of regulations and standards (e.g., [16,42,46]). Second, regarding pragmatic knowledge contributions, *methods, tools, and frameworks* facilitate AI auditing for researchers, developers, auditors, and AI system users (e.g., [10,49]). This knowledge contribution is tied closely to design science studies, which produce a concrete artifact. By methods, we mean systematic procedures, e.g., algorithms and practices, while tools are concrete instruments to perform tasks, and frameworks are more high-level structures that help organize activities. Third, in terms of normative knowledge contributions, *awareness and empowerment* enable AI system users and the wider public to understand and control AI systems (e.g., [111,122]). Each type of knowledge contribution (cognitive, pragmatic, and normative) represents one motivating driver for ethics-based AI auditing. Particular studies may serve multiple purposes and stakeholders simultaneously.

The synthesizing framework presented in Fig. 2 summarizes the knowledge contributions to stakeholders and the ethical principles in the ethics-based AI auditing literature. The center of the figure shows the most prominent knowledge contributions for the AI auditing stakeholder groups, and the outer ring lists the ethical principles that delimit the ethics-based AI auditing literature. The conceptualizations (and eventual operationalization) of these principles shape how ethics-based AI auditing is conducted and the extent to which it can tackle the risks and potential impacts of AI systems. In line with our research questions, our analysis focuses on conceptualizations of principles, knowledge contributions, and stakeholder groups. We discuss the operationalization of AI auditing as one point in the future research agenda below (Section 5.5).

The framework elucidates the multiple dimensions of ethics-based AI auditing as well as the synergistic and potentially conflicting goals of the AI auditing literature. The framework, thus, provides an overall map of principles, stakeholders, and intended contributions for positioning individual studies and making deliberate decisions about their scope. For example, an AI auditing study may provide cognitive knowledge contributions to regulators regarding justice and fairness while excluding normative and pragmatic contributions and other stakeholders and principles. Another study could provide tools (pragmatic knowledge contributions) to AI system developers to enhance trustworthiness and transparency. A third study may seek to present a comprehensive auditing methodology (pragmatic knowledge contribution) to AI system users, covering fairness, transparency, non-maleficence, responsibility, and privacy. Each of these hypothetical studies covers a different area of the overall space of ethics-based AI auditing.

Although the framework gives a map of ethics-based AI auditing, it is a synthesis of the present state of the literature rather than a normative statement of the ideal state of things. Hence, based on the patterns in our analysis, certain gaps can be identified in the current AI auditing literature. First, the arrows in the framework illustrate that different stakeholders tend to receive different knowledge contributions. For example, cognitive knowledge contributions are provided to regulators, researchers, and developers, while pragmatic knowledge contributions are provided to a wide range of stakeholders. However, normative contributions (awareness and empowerment) are also necessary for

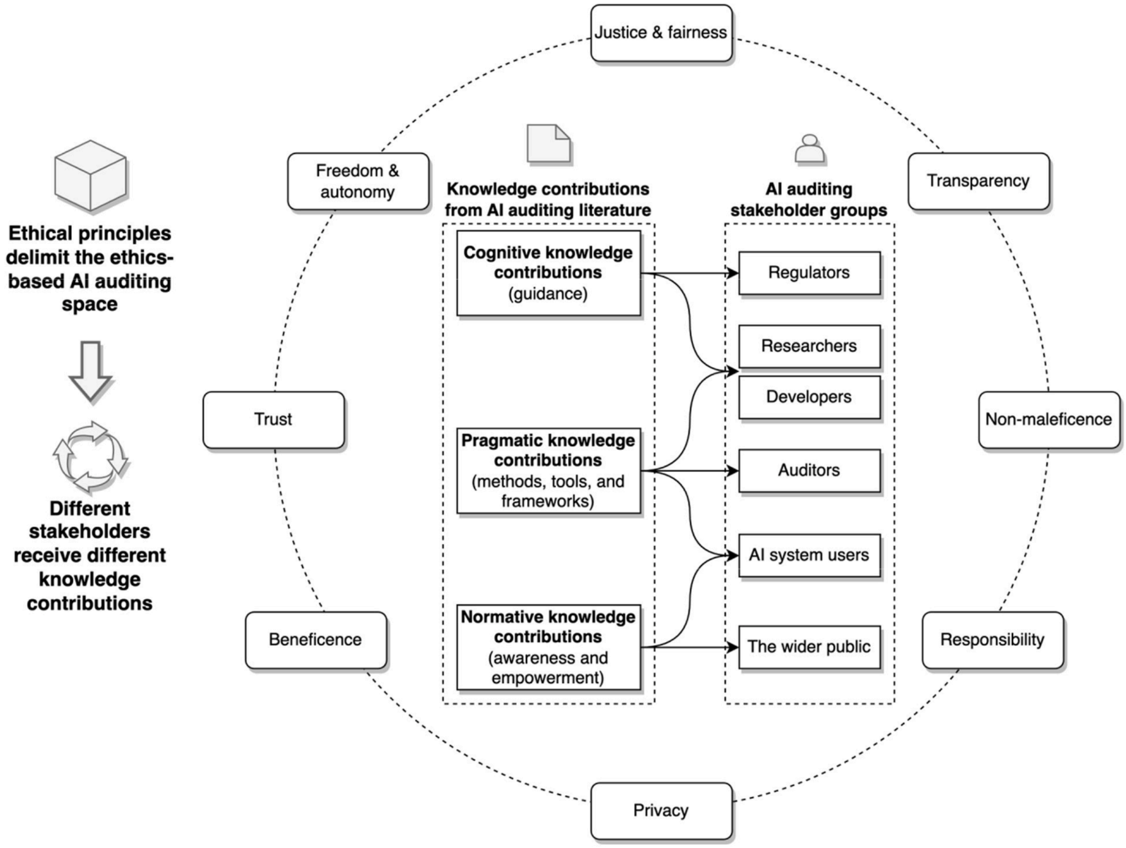
Ethical principles delimit the ethics-based AI auditing space
Different stakeholders receive different knowledge contributions
Justice & fairness
Freedom & autonomy
Transparency
Trust
Non-maleficence
Beneficence
Responsibility
Privacy
Knowledge contributions from AI auditing literature
Cognitive knowledge contributions (guidance)
Pragmatic knowledge contributions (methods, tools, and frameworks)
Normative knowledge contributions (awareness and empowerment)
AI auditing stakeholder groups
Regulators
Researchers
Developers
Auditors
AI system users
The wider public

Fig. 2. Framework synthesizing the ethical principles, stakeholders, and knowledge contributions in the AI auditing literature.

developers, researchers, and regulators dealing with AI systems, not only the users and the wider public. The current literature seems to assume the ethical awareness and agency of these former stakeholder groups. Still, the messiness of AI accountability topics with multiple issues, forums, and audiences (e.g., [148]) indicates that ethical empowerment is also crucial for researchers, developers, and regulators. Second, pragmatic knowledge contributions (tools and methods) are predominant in the literature. At present, the ethics-based AI auditing literature is focused on technical tools for relatively narrow issues at the expense of conceptualizing the AI auditing space, making sense of auditing processes, and problematizing the 'ethics' in ethics-based AI auditing. Third, it is not evident that the delimiting ethical principles are exhaustive. For example, principles such as environmental sustainability and human dignity could be envisioned as additional ethical bases for AI auditing, and they could give rise to new auditing practices and literature streams. Fourth, while the reviewed studies featured several stakeholders, some key stakeholder groups were not present in the reviewed literature. Organizational managers were not among the stakeholders the ethics-based AI auditing studies targeted. In some cases, auditing was conducted for organizational purposes (e.g., [119]), but AI auditing primarily aimed to benefit individuals and groups outside of organizations. The lack of a managerial audience is surprising because managers are expected to be concerned with whether an organization's AI systems perform according to requirements.

The lack of connection to a managerial audience indicates more general questions about the current ethics-based AI auditing literature. For example, in AI ethics and ethics-based AI auditing discourses, "ethics" tends to be framed somewhat narrowly, most often in terms of bias, as evidenced by the number of studies discussing bias in the "justice and fairness" category. This tendency is understandable because bias lends itself to formal operationalization and technical mitigation methods. Simultaneously, potential ethical issues around AI systems are much broader and are linked to potential job loss, autonomous warfare, and human actors' role in increasingly automated work practices. The ethics-based AI auditing literature seems to primarily focus on issues amenable to engineering-based solutions rather than addressing the more wide-ranging societal and accountability questions around autonomous systems. Whether a demand for broader ethics-based AI auditing exists is an essential question for subsequent research.

Investors and funders of organizations also represent a neglected stakeholder group that may be highly interested in audit information. In particular, the trend of sustainable investment could be extended to include AI ethics considerations, but this development remains in its infancy [17,93]. Moreover, even though some studies targeted users (17) and regulators (16), these are a small subset of the 93 analyzed articles. Based on our synthesis, the ethics-based AI auditing literature is not yet a mature field, and its conceptualization and operationalization, especially from the organizational and social perspectives, remain incipient.

5.3. Implications for is research and practice

With the syntheses on the conceptualizations of focal ethical concepts and the knowledge contributions to stakeholders, this study contributes to two literature streams: the IS literature on AI in organizations (e.g., [11,12,79]) and the literature on AI system stakeholders (e.g., [75,87]). The previous IS literature on AI within organizations has established that the novel characteristics of AI technologies (autonomy, learning, and inscrutability) and the continuously advancing nature of AI capabilities require continuously emerging management efforts [12]. With careful coordination, AI systems can facilitate effective organizational learning in turbulent environments [140]. Still, the inherent uncertainty of knowledge in knowledge work accentuates the difficulty of training AI systems with experts' know-how in addition to know-what [78]. When introduced to organizational contexts, AI systems bring significant risks of unfairness, bias, and discriminatory decisions [31,44], which need to be tackled with appropriate oversight and accountability [143]. We also know from previous literature that AI governance is a complex set of organizational practices and processes connected to an organization's governance system [105,110], where top management and board characteristics can facilitate setting strategic directions for AI within organizations [80].

We extend the IS literature on AI in organizations in two ways. First, we survey the entire set of ethical principles that need to be considered in AI auditing rather than focusing on tackling single issues such as fairness. The organizational AI literature can be divided into two areas: a narrow literature stream focusing on specific issues such as fairness or accountability [31,44,52] and a broad literature stream discussing the entire system of governance issues [13,110,127]. Although the narrow body of research is more mature, our study contributes in particular to the broad, more emerging body of research. The comprehensive approach to ethical principles underscores the complexity of AI auditing as a part of AI management. Effective AI auditing requires simultaneous consideration of numerous ethical principles, which may require different approaches, frameworks, and tools. Second, with the matrix in Table 7, we elucidate the goals of AI auditing (the conceptualizations of ethical principles) and the different perspectives on AI auditing depending on the chosen socio-technical (the social/technical dimension) and audit approach (the process/outcome dimension) emphases. From a practical standpoint, this distinction between different conceptualizations underscores organizations' need to understand which kinds of conceptualizations (and subsequent operationalizations) support effective AI management and AI auditing in their contexts.

Our second contribution focuses on the literature on AI system stakeholders (e.g., [75,87]). From the previous IS literature, we know that managing AI requires considering multiple stakeholders in terms of their synergistic or conflictual expectations and actions [94,139]. More specifically, explainable AI requires carefully considering the target audiences of explanations [75,87]. In addition, different types of algorithmic accountability relations are formed between stakeholder groups such as institutions, organizations, and users [52].

With the framework in Fig. 2, we complement and go beyond the AI system stakeholder literature by clarifying what knowledge contributions stakeholders are expected to receive from the AI auditing literature and, thus, what beneficial functions the AI auditing literature has for stakeholders. By doing so, we shed light on the role of AI auditing research in the multi-stakeholder ecosystem of AI management (e.g., [139]). This helps stakeholders, such as AI system developers and users, understand what they can seek in AI auditing research, and, conversely, it helps researchers illuminate their core message to stakeholders, strengthening the research impact and the science-society interface.

5.4. Limitations

This study's main limitations are threefold. First, the identification of ethics-based studies and ethical principles is based on existing publications (AI [1,59]). Therefore, this review focused only on the ethics-based facet of AI auditing as it exists in the current literature. The studies were screened and analyzed based on the existing principles, and the study did not aim to recognize new or unidentified areas. Future studies could add other sectors to the analysis, identify principles that have not been considered yet, or use alternative ways of screening the literature using particular principles. Also, the inclusion criterion for published literature considered only academic publications. Future studies could add sources from gray literature while acknowledging that this would make the data set more heterogeneous.

Second, with its focus on AI ethics principles and stakeholders, this review did not consider the details of the auditing process nor the practical implementation or position of internal auditing within organizations. The perspective was ethics-based, which may divert attention from other perspectives, such as those investigated in the risk and controls matrix by KPMG [68]. The matrix identifies supplier management,

Table 9
Future research agenda for ethics-based AI auditing.

Research topic	Future research agenda	Potential research issues
Cognitive knowledge contributions		
Technical and legal perspectives on AI auditing	Review AI auditing literature through technical and legal perspectives Compare AI auditing to other types of IT auditing	The implications of technical developments and tools on AI auditing Technical requirements for AI systems posed by AI auditing The implications of existing and emerging legislation on AI auditing
Incorporate new generations of AI technologies	Investigate ethical considerations related to the use of generative and conversational AI Develop guidelines and frameworks for the ethical use of generative AI	Address potential biases and unintended consequences of conversational AI
Pragmatic knowledge contributions		
Frameworks for ethics-based AI auditing	Practical solutions and guidelines for implementing AI auditing into practice	Reasons and remedies for the lack of practical solutions of AI auditing Determining what ethical principles should guide AI decision making
Auditing process and implementation	Study business-process-oriented AI auditing perspectives and their connections to ethics-based auditing Investigate successful cases of AI auditing in practice and explore best practices	Exploring concrete solutions for conducting AI audits and implementing AI auditing into practice
Normative knowledge contributions		
Broadening ethics-based AI auditing	Identify other sectors to analyze, consider gray literature sources, and recognize new principles	Investigating why certain principles and stakeholder groups are omitted in the current literature Exploring e.g. sustainability or solidarity-based frameworks
Human-centric and socio-technical AI auditing	Develop socio-technical methods and frameworks to improve interpretability, bias mitigation, data protection, holding AI systems accountable and ensuring fairness	Understanding the roles of people, software, and human-computer interactions in AI auditing Strengthening human-centricity and accountability to stakeholders

business processes, and other business perspectives, different from the ethics-based AI auditing focus adopted in this paper. These business-process-oriented AI auditing perspectives and their connections to ethics-based auditing could be studied in future research. More broadly, future research could also review the AI auditing literature spectrum through aspects other than ethics, e.g., technical or legal perspectives not discussed in this paper.

Third, our study did not consider AI standards' relevance for ethics-based AI auditing. AI standards are currently emerging (cf. [22]); thus, considering standards in addition to principles would add considerably to the complexity level of research. Studies on AI standards' relevance for auditing are warranted in future research.

5.5. Future research agenda

Based on our SLR findings, we present a future research agenda for ethics-based auditing research (Table 9). The research agenda is divided into the three types of knowledge contributions discussed previously: cognitive, pragmatic, and normative knowledge contributions. The knowledge contributions represent AI auditing research streams to which future studies can contribute. Each topic in Table 9 is followed by a statement of the broad future research agenda and more concrete potential research issues.

Cognitive knowledge contributions can be furthered by reviewing the ethics-based or broader AI auditing literature from technical and legal perspectives, complementing the ethics-based view in this paper. In addition, future studies could compare AI auditing to other types of IT auditing to clarify the novel aspects of AI auditing. New generations of AI technologies, such as generative AI and the large language models [60] used in the ChatGPT chatbot, for example, should also be addressed by investigating their specific ethical considerations and developing the necessary guidelines and frameworks for auditing new generations of AI technologies.

In turn, pragmatic knowledge contributions can be made by developing robust frameworks for ethics-based AI auditing with well-defined principles and practical guidelines. The auditing process and implementation should also be elucidated from a business process perspective, which could eventually embed AI auditing into the core business processes of consultancies providing auditing services as well as the development processes of companies developing AI.

For normative knowledge contributions, the findings on the most prevalent ethical principles imply that future research should broaden ethics-based AI auditing and focus on other sectors, gray literature sources, and the principles not included in the literature. Why are dignity, sustainability, and solidarity not discussed in the current ethics-based AI auditing literature? In future auditing approaches, would it be possible to integrate more collective sustainability or solidarity-based frameworks in addition to the predominant focus on individuals? Similarly, the underrepresented stakeholder groups, including organizational managers and investors, represent promising avenues for future research to understand ethics-based AI auditing comprehensively. Moreover, the socio-technical interfaces in AI auditing and the roles of people, software, and human-computer interactions are promising future research directions. Although many questions remain, this systematic review paves the way for ongoing empirical research into the thus-far largely uncharted area of ethics-based AI auditing in the pursuit of more socially responsible algorithmic systems.

CRediT authorship contribution statement

Joakim Laine: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matti Minkkinen:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Matti Mäntymäki:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors have no competing interests to declare.

Acknowledgment

The research reported here was conducted within the Artificial Intelligence Governance and Auditing (AIGA) project, funded by Business Finland's AI Business program.

Appendix A. Studies included in the review, their methodological approaches, and main outputs

Study	Article	Method	Main output
P1	[118]	Design science	The SMACTR framework for algorithmic auditing that supports AI system development end-to-end
P2	[76]	Design science	An ethical AI algorithm audit framework
P3	[83]	Conceptual and design science	The new concept of algorithm 'legibility' in order to combine transparency and comprehensibility
P4	[4]	Design science	A method of auditing outputs for social biases
P5	[64]	Experiment	An empirical investigation of the "SUBGROUP" algorithm on four data sets
P6	[117]	Case study	Outlines the audit design and structured disclosure procedure used in the Gender Shades study, presents new performance metrics on the Pilot Parliaments Benchmark (PPB) and provides performance results on PPB
P7	[65]	Design science and case study	A framework of multi-accuracy auditing and post-processing to improve predictor accuracy across identifiable subgroups.
P8	[32]	Design science and experiment	An approach whereby individual subjects on whom automated decisions are made can elicit in a collaborative and privacy-preserving manner a rule-based approximation of the model underlying the decision algorithm
P9	[141]	Design science	A reinforcement learning based framework for algorithmic bias detection in ML powered autonomous software systems and a control loop which allows fairness check in decisions of autonomous software systems
P10	[84]	Conceptual	A multi-agent system architecture
P11	[19]	Design science	"FAIRVIS," a visual analytics system for discovering intersectional bias in machine learning models
P12	[135]	Conceptual	The concept of decision provenance to provide information exposing decision pipelines: chains of inputs to, the nature of, and the flow-on effects from the decisions and actions taken throughout systems
P13	[119]	Design science and case study	"CelebSET," an audit process for products employing facial processing technology
P14	[42]	Focus groups and a digital ethnography	An algorithmic audit of "REM!X," a personalized well-being recommendation app
P15	[34]	Experiment	Audit of "ArcFace," a state-of-the-art, open-source face recognition system
P16	[50]	Survey-based experiment	A study on perceptions of fairness in machine learning models
P17	[111]	Design science	A complete overview of bias in word embeddings
P18	[14]	Design science	"FlipTest," a black-box technique for uncovering discrimination in classifiers
P19	[54]	Design science	A stylized model and fairness requirements that match the intuitive fairness desiderata
P20	[62]	Design science	Algorithmic Equity Toolkit
P21	[26]	Simulation	An extensible open-source software framework for implementing fairness-focused simulation studies and further reproducible research
P22	[137]	Experiment	Audit an existing cyberbullying algorithm using Twitter data for disparity in detection performance based on the network centrality of the potential victim
P23	[5]	Experiment	Factors influencing the perception of "fairness"
P24	[125]	Design science	A novel metric for auditing group fairness in ranked lists
P25	[21]	Case study and experiment	Gender-based inequalities in the context of resume search engines
P26	[86]	Design science and case study	A framework for internally auditing online services
P27	[72]	Design science and case study	A framework to measure biases in Twitter's search results
P28	[122]	Survey	Results of a mixed-methods algorithm audit of partisan audience bias and personalization within Google Search
P36	[16]	Design science	An auditing framework to guide ethical assessment of an algorithm
P37	[144]	Design science and exploratory research	A prototype that performs audits on social networks
P38	[133]	Conceptual	15 governance recommendations for creating reliable, safe, and trustworthy human-centered AI systems across team, organizational, and industry levels
P39	[6]	Experiment	A study on the extent to which image tagging algorithms mimic the phenomenon of learning social stereotypes through observation
P40	[126]	Qualitative analysis	A study on how race and gender are defined and annotated in image databases used for facial analysis
P41	[47]	Case study	A framework for applying algorithmic accountability and ethical principles to ecological forecasting models
P42	[58]	Experiment	Statistical differences in MTurk annotators' performances when different modalities of information are provided and discuss the patterns of harm that arise from crowdsourced human demographic prediction
P29	[18]	Experiment	An approach to evaluate bias in automated facial analysis algorithms and data sets
P30	[69]	Conceptual	A new technological toolkit for automated decisions and standards of legal fairness
P31	[124]	Conceptual	Outlined five idealized audit designs for empirical research projects investigating algorithms
P32	[142]	Design science	A transparent model distillation approach to detect bias in black box risk scoring models
P33	[136]	Conceptual	Discussion paper about responsibility in ML, focusing on the importance of transparency and control within ML workflows and their societal impacts
P34	[121]	Conceptual	Investigation of legal liability in ML
P35	[46]	Conceptual	Investigation about how EU GDPR addresses discrimination
P43	[40]	Conceptual	AI4People—An Ethical Framework for a Good AI Society
P44	[96]	Conceptual	A critical assessment of the strategies and recommendations proposed by current AI Ethics initiatives
P45	[95]	Conceptual	A conceptual framework aiming to inform future ethical inquiry, development, and governance of algorithms
P46	[108]	Case study	A study showing that a widely used algorithm, in a context of health systems, exhibits significant racial bias
P47	[73]	Mixed methods	A set of descriptive tags for all images in the Chicago Face Database using the six tagging APIs
P48	[10]	Design science	AI Fairness 360 Toolkit, a new open-source Python toolkit for algorithmic fairness aims to bridge fairness research in algorithms with industrial application and offers a framework for researchers to share and evaluate their algorithms
P49	[49]	Conceptual	Ground conceptualizations of race for fairness research
P50	[85]	Survey	A comprehensive survey of biases in AI systems and a categorization of fairness definitions
P51	[134]	Experiment	A between-subjects user study to examine various explanation styles' effects on users' fairness perceptions
P52	[129]	Case study	The audit revealed discrepancies in algorithm documentation and practice, subjective variables in the model, and oversight in the ethical use of personal characteristics for employment predictions.
P53	[93]	Interview	Indicates that AI responsibility is not yet a standard part of ESG investment analyses, highlighting the need for standardized AI governance metrics.
P54	[112]	Conceptual	A socio-computational interrogation of Google searching by image algorithm
P55	[77]	Conceptual	Psychological audits as a standardized approach for evaluating fairness and bias

(*continued on next page*)

(*continued*)

Study	Article	Method	Main output
P56	[123]	Case study	A model assessing transparency and accountability in Brazilian digital public services
P57	[3]	Conceptual	The mechanisms for the next steps that can help the public assess the trustworthiness of AI developers
P58	[36]	Design science	Systematic, principled, and general approach to audit ML models
P59	[151]	Design science	A novel auditing protocol AP-Aml, for image privacy and efficiency in ambient intelligence systems
P60	[132]	Conceptual and case study	The concept of "everyday algorithm auditing" where users in their daily use of digital platforms detect and report algorithmic biases and harmful behaviors
P61	[146]	Conceptual	History of tensions that have shaped the development of social science audits
P62	[109]	Design science	“FairLens,” a methodology for discovering and explaining biases
P63	[102]	Conceptual	Ethics-based auditing (EBA) as a governance mechanism
P64	[82]	Conceptual	An analysis of the regulatory content of 16 guideline documents about the use of AI in the public sector
P65	[145]	Conceptual	The PLEAD project demonstrates how computable explanations can enhance GDPR compliance and empower both data controllers and subjects
P66	[8]	Experiment	Implement a sock-puppet audit to audit black-box social media systems
P67	[150]	Design science	An explorative model building system “FairRover” for responsible fair model building
P68	[55]	Design science	A new methodology for black-box auditing of algorithms for discrimination in the delivery of job ads
P69	[51]	Conceptual	A combined Acceptance Test-Driven Development (ATDD) and Assurance Cases approach to assure and articulate the fairness of algorithmic decision-making systems
P70	[149]	Case study	A framework for algorithmic auditing
P71	[100]	Conceptual	Outlines of the conditions under which ethics-based auditing procedures can be feasible and effective in practice
P72	[101]	Conceptual	A comparison of the European AI Act’s enforcement mechanisms with AI auditing literature and offers amendments for clarity and improved regulatory practices
P73	[37]	Case study	An explanatory case study aimed at examining bias
P74	[138]	Design science	A matrix for auditing algorithmic decision-making systems (ADSs)
P75	[25]	Conceptual	A systematic analysis of the machine learning pipeline
P76	[71]	Case study	An audit of a widely used OpenCV algorithm for pupil detection
P77	[27]	Conceptual	Contextualize anti-blackness in the design, development, and deployment of AI systems
P78	[2]	Design science	Data, auditing, monitoring, and output, i.e., DAMO taxonomy
P79	[63]	Survey	Analyzes fairness, explainability, accountability, reliability, and acceptance requirements of trustworthy AI systems
P80	[147]	Conceptual	Recommendations for implementing data provenance in AI systems to mitigate bias and promote responsible AI
P81	[98]	Survey	A framework intended to share knowledge of and experiences with XAI design and evaluation methods
P82	[132]	Case study	Propose and explore the concept of everyday algorithm auditing
P83	[20]	Design science and interview	“Deblinder,” a visual analytics system for synthesizing failure reports
P84	[120]	Interview	A framework for analyzing how organizational culture and structure impact the effectiveness of responsible AI initiatives in practice
P85	[88]	Mixed methods	Persistent underrepresentation of women and people of color in image search results for occupations and demonstrating that such representations can influence users’ perceptions
P86	[114]	Design science	A fairness audit framework that assesses the fairness of ML algorithms while addressing potential security issues such as data privacy, model secrecy, and trustworthiness
P87	[116]	Case study	Reports on the impacts of using publicly available visualization tools used in HCI practice
P88	[24]	Conceptual	Reviewability as a framework that involves breaking down the automated and algorithmic decision making into technical and organizational elements to provide a systematic framework for determining the contextually appropriate record-keeping mechanisms to facilitate meaningful review
P89	[29]	Interviews and workshops	A process model that captures the dynamics of and influences on users’ search and sensemaking behaviors
P90	[66]	Conceptual	A model of public trust in AI that provides a theoretical scaffolding for trusted AI research
P91	[115]	Participatory workshops and interviews	A stakeholder-centered design ideas for solutions to mitigate tensions surrounding AI in human resource management
P92	[70]	Literature review	A systematic analysis of traceability in AI governance, mapping requirements to available technologies and identifying existing gaps for accountability.
P93	[[45]	Conceptual	The envisioned INFER framework that aims to increase trust in machine-generated recommendations

Appendix B. Ethical principles and methods in the reviewed studies

	Design science	Other empirical	Conceptual	Total no. of studies	Included codes
Justice & fairness	P1,P7,P9,P10,P11,P13,P17, P18,P19,P20,P21,P24,P27, P32,P36,P48,P58,P60,P62, P66,P67,P68,P70,P74,P83, P86,P88,P90,P93	P4,P5,P6,P8,P14,P15,P16,P22, P23,P25,P26,P28,P29,P39,P41, P42,P46,P47,P50,P51,P53 P56, P73,P76,P79,P80,P81,P82,P84, P85,P87,P89,P91	P2,P3 P12,P30,P31,P34, P35,P38,P43,P45,P49, P54,P55,P61,P63,P69, P71,P72,P75,P77,P92	83	Justice,fairness,consistency,inclusion, equality,equity,bias,discrimination, diversity,plurality,accessibility,reversibility, remedy,redress,challenge,access,distribution
Transparency	P1,P13,P18,P20,P32,P36, P48,P60,P62,P70,P74,P88, P90,P93	P6,P8,P14,P23,P25,P29,P39,P41, P42,P47,P50,P51,P52,P53 P56, P57,P64,P73,P79,P80,P81,P84, P87,P91	P3,P12,P30,P31,P33,P34, P35,P38,P43,P45,P55, P63,P65,P71,P72,P92	54	Transparency,explainability,explicability, understandability,interpretability, communication,disclosure,showing
Non-maleficence	P1,P7,P13,P20,P21,P36,P59, P60,P70,P78,P88,P90	P6,P14,P37,P42,P50,P53,P57, P64,P79,P82,P85,P89,P91	P2,P3,P12,P30,P31,P33, P34,P35,P38,P43,P44, P55,P61,P63,P71,P72, P75	41	Non-maleficence,security,safety,harm, protection,precaution,prevention,integrity, non-subversion
Responsibility	P1,P13,P20,P67,P68,P74, P88,P90	P6,P14,P16,P23,P29,P41,P42, P52,P53,P56,P64,P79,P80,P81, P84,P87	P12,P30,P31,P33,P34, P35,P38,P43,P44,P54, P63,P65,P71,P72,P92	39	Responsibility,accountability,liability,acting with integrity

(*continued on next page*)

(*continued*)

	Design science	Other empirical	Conceptual	Total no. of studies	Included codes
Privacy	P1,P13,P59,P68,P78,P86,P88	P6,P8,P53,P56,P57,P79,P80,P81,P84,P91	P2,P3,P12,P30,P33,P55,P63,P75	25	Privacy,personal or private information
Trust	P13,P62,P70,P86,P90,P93	P15,P51,P57,P73,P79,P80,P81,P91	P3,P30,P34,P43,P55,P63,P71,P92	22	Freedom,autonomy,consent,choice,self-determination,liberty,empowerment,trust
Beneficence	P10,P48	P8,P53,P56,P64,P73,P79,P80,P81,P82	P3,P12,P30,P34,P38,P43,P63,P71,P72	20	Benefits,beneficence,well-being,peace,social good,common good
Freedom and autonomy	P36,P60,	P16,P28,P64,P79	P12,P30,P38,P43,P63,P71,P72	13	Freedom,autonomy,consent,choice,self-determination,liberty,empowerment
Sustainability	(not discussed in the reviewed studies)				Sustainability,environment,energy,resources
Dignity	(not discussed in the reviewed studies)				Dignity
Solidarity	(not discussed in the reviewed studies)				Solidarity,social security,cohesion

References

[1] AI HLEG, Ethics Guidelines for Trustworthy AI. Independent High-level Expert Group on Artificial Intelligence set by European Commission, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[2] M. Akmal, B. Syangtan, A. Alchouemi, Enhancing the security of data in cloud computing environments Using Remote Data Auditing, in: 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), 2021, pp. 1–10, https://doi.org/10.1109/CITISIA53721.2021.9719899.

[3] S. Avin, H. Belfield, M. Brundage, G. Krueger, J. Wang, A. Weller, M. Anderljung, I. Krawczuk, D. Krueger, J. Lebensold, T. Maharaj, N. Zilberman, Filling gaps in trustworthy development of AI, Science (1979) 374 (6573) (2021) 1327–1329, https://doi.org/10.1126/science.abi7176.

[4] P. Barlas, K. Kyriakou, S. Kleanthous, J. Otterbacher, Social B(eye)as: human and machine descriptions of people images, Proceed. Thirt. Int. AAAI Conferen. Web Soc. Media 13 (1) (2019) 583–591.

[5] P. Barlas, S. Kleanthous, K. Kyriakou, J. Otterbacher, What Makes an Image Tagger Fair?, in: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, 2019, pp. 95–103.

[6] P. Barlas, K. Kyriakou, O. Guest, S. Kleanthous, J. Otterbacher, To "See" is to stereotype: image tagging algorithms, gender recognition, and the accuracy-fairness trade-off, Proc. ACM. Hum. Comput. Interact. 4 (3) (2020) 1–31.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities, and challenges toward responsible AI, Inform. Fus. 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.

[8] N. Bartley, A. Abeliuk, E. Ferrara, K. Lerman, Auditing algorithmic bias on Twitter, in: 13th ACM Web Science Conference 2021, 2021, pp. 65–73, https://doi.org/10.1145/3447535.3462491.

[9] F. Batarseh, A. Freeman, C.-H. Huang, A survey on artificial intelligence assurance, J. Big. Data 8 (1) (2021) 60.

[10] Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K., & Zhang, Y. (2018). AI Fairness 360: an Extensible Toolkit For Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv preprint arXiv:1810.01943.

[11] H. Benbya, T.H. Davenport, S. Pachidi, Artificial intelligence in organizations: current state and future opportunities, MIS Q. Execut. 19 (4) (2020).

[12] N. Berente, B. Gu, J. Recker, R. Santhanam, Managing artificial intelligence, MIS Q. 45 (3) (2021) 1433–1450.

[13] T. Birkstedt, M. Minkkinen, A. Tandon, M. Mäntymäki, AI Governance: themes, Knowledge Gaps, and Future Agendas, Internet Research (2023).

[14] E. Black, S. Yeom, M. Fredrikson, FlipTest: fairness testing via optimal transport, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 111–121.

[15] W.M. Bramer, M.L. Rethlefsen, J. Kleijnen, et al., Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study, Syst. Rev. 6 (2017) 245, https://doi.org/10.1186/s13643-017-0644-y, 2017.

[16] S. Brown, J. Davidovic, A. Hasan, The algorithm audit: scoring the algorithms that score us, Big. Data Soc. 8 (1) (2021).

[17] J. Brusseau, AI human impact: toward a model for ethical investing in AI-intensive companies, J. Sustain. Finance Invest. (2021) 1–28.

[18] J. Buolamwini, T. Gebru, Gender shades: intersectional accuracy disparities in commercial gender classification, in: *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, 2018, pp. 77–91.

[19] A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, D.H. Chau, FAIRVIS: visual analytics for discovering intersectional bias in machine learning, in: IEEE Conference on Visual Analytics Science and Technology, 2019, pp. 46–56.

[20] Á.A. Cabrera, A.J. Druck, J.I. Hong, A. Perer, Discovering and validating AI errors with crowdsourced failure reports, Proc. ACM. Hum. Comput. Interact. 5 (CSCW2) (2021) 1–22, https://doi.org/10.1145/3479569.

[21] L. Chen, A. Hannak, R. Ma, C. Wilson, Investigating the impact of gender on rank in resume search engines, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–14.

[22] P. Cihon, Standards For AI Governance: International Standards to Enable Global Coordination in AI Research & Development, Future of Humanity Institute, Oxford, 2019.

[23] L. Clarke, AI auditing is the Next Big Thing. But will it Ensure Ethical Algorithms?, 2021.

[24] J. Cobbe, M.S.A. Lee, J. Singh, Reviewable automated decision-making: a framework for accountable algorithmic systems, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 598–609, https://doi.org/10.1145/3442188.3445921.

[25] J. DeHart, C. Xu, C. Grant, L. Egede, Proposing an interactive audit pipeline for visual privacy research, in: 2021 IEEE International Conference on Big Data, 2021, pp. 1249–1255, https://doi.org/10.1109/BigData52589.2021.9671478.

[26] A. D'Amour, P. Baljekar, H. Srinivasan, D. Sculley, J. Atwood, Y. Halpern, Fairness is not static: deeper understanding of long-term fairness via simulation studies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 525–534.

[27] C.L. Dancy, P.K. Saucier, AI and blackness: toward moving beyond bias and representation, IEEE Trans. Technol. Soc. 3 (1) (2022) 31–40, https://doi.org/10.1109/TTS.2021.3125998.

[28] T. Davenport, A. Guha, D. Grewal, T. Bressgott, How artificial intelligence will change the future of marketing, J. Acad. Market. Sci. 48 (1) (2020) 24–42.

[29] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, M. Eslami, Toward User-Driven Algorithm Auditing: investigating users' strategies for uncovering harmful algorithmic behavior, in: CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–19, https://doi.org/10.1145/3491102.3517441.

[30] V. Dignum, Responsible Artificial Intelligence – How to Develop and Use AI in a Responsible Way, Springer, 2019.

[31] M. Dolata, S. Feuerriegel, G. Schwabe, A sociotechnical view of algorithmic fairness, Inform. Syst. J. 32 (4) (2022) 754–818, https://doi.org/10.1111/isj.12370.

[32] J. Domingo-Ferrer, C. Perez-Sola, A. Blanco-Justicia, Collaborative explanation of deep models with limited interaction for trade secret and privacy preservation, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 501–507.

[33] B. Dowden, Fallacies | Internet Encyclopedia of Philosophy, 2022. https://iep.utm.edu/fallacy.

[34] C. Dulhanty, A. Wong, Investigating the impact of inclusion in face recognition training data on individual face identification, in: 2020 AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 244–250.

[35] A. Dutta, R. Roy, P. Seetharaman, An assimilation maturity model for IT governance and auditing, Inform. Manag. 59 (1) (2022) 103569, https://doi.org/10.1016/j.im.2021.103569.

[36] F.-E. Eid, H.A. Elmarakeby, Y.A. Chan, N. Fornelos, M. ElHefnawi, E.M. Van Allen, L.S. Heath, K. Lage, Systematic auditing is essential to debiasing machine learning in biology, Commun. Biol. 4 (1) (2021) 183, https://doi.org/10.1038/s42003-021-01674-5.

[37] Y. Ennali, T. Engers, Data-driven AI development: an integrated and iterative bias mitigation approach, in: CEUR Workshop Proceedings, 3rd EXplainable AI in Law Workshop, 2021.

[38] European Commission, in: Proposal for A Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts Com/2021/206 Final, 2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

[39] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, M. Jirotka, H. Johnson, C. LaPointe, A.J. Llorens, A. K. Mackworth, C. Maple, S.E. Pálsson, F. Pasquale, A. Winfield, Z.K. Yeong, Governing AI safety through independent audits, Nat. Mach. Intell. 3 (7) (2021) 566–571, https://doi.org/10.1038/s42256-021-00370-7.

[40] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—an ethical framework for a good AI Society: opportunities, risks, principles, and recommendations, Mind. Mach. (2018) 689–707.

[41] L. Floridi, J. Cowls, A unified framework of five principles for AI in society, Harv. Data Sci. Rev. 1 (1) (2019), https://doi.org/10.1162/99608f92.8cd550d1.

[42] G. Galdon Clavell, M. Zamorano, C. Castillo, O. Smith, A. Matic, Auditing algorithms: on lessons learned and the risks of data minimization, in: Proceedings of 2020 ACM AI, Ethics, and Society Conference, 2020, pp. 265–271.

[43] U. Gasser, V.A.F. Almeida, A layered model for AI governance, IEEE Internet. Comput. 21 (2017) 58–62.

[44] M. Ghasemaghaei, N. Kordzadeh, Understanding how algorithmic injustice leads to making discriminatory decisions: an obedience to authority perspective, Inform. Manag. (2024), https://doi.org/10.1016/j.im.2024.103921.

[45] G. Giannopoulos, G. Papastefanatos, D. Sacharidis, K. Stefanidis, Interactivity, fairness, and explanations in recommendations, in: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation, and Personalization, 2021, pp. 157–161, https://doi.org/10.1145/3450614.3462238.

[46] B. Goodman, A step toward accountable algorithms? Algorithmic discrimination and the European Union general data protection, in: 29th Conference on Neural Information Processing Systems, 2016.

[47] I. Grasso, D. Russell, A. Matthews, J. Matthews, N. Record, Applying algorithmic accountability frameworks with domain-specific codes of ethics: a case study in ecosystem forecasting for shellfish toxicity in the Gulf of Maine, in: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, 2020, pp. 83–91.

[48] T. Hagendorff, The ethics of AI ethics: an evaluation of guidelines, Mind. Mach. (Dordr.) 30 (1) (2020) 99–120, https://doi.org/10.1007/s11023-020-09517-8.

[49] A. Hanna, E. Denton, A. Smart, J. Smith-Loud, Toward a critical race methodology in algorithmic fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2019, pp. 501–512.

[50] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, B. Ur, An empirical study on the perceived fairness of realistic, imperfect machine learning models, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 392–402.

[51] M.P. Hauer, R. Adler, K. Zweig, Assuring fairness of algorithmic decision making, in: 2021 IEEE International Conference on Software Testing, Verification, and Validation Workshops (ICSTW), 2021, pp. 110–113, https://doi.org/10.1109/ICSTW52544.2021.00029.

[52] D. Horneber, S. Laumer, Algorithmic Accountability, Business & Information Systems Engineering 65 (6) (2023) 723–730, https://doi.org/10.1007/s12599-023-00817-8.

[53] X. Hu, R. Rousseau, J. Chen, On the definition of forward and backward generations, J. Informetr. 5 (1) (2011) 27–36.

[54] C. Ilvento, M. Jagadeesan, S. Chawla, Multi-Category Fairness in Sponsored Search Auctions, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 348–358.

[55] B. Imana, A. Korolova, J. Heidemann, Auditing for discrimination in algorithms delivering job ads, in: Proceedings of the Web Conference 2021, 2021, pp. 3767–3778, https://doi.org/10.1145/3442381.3450077.

[56] Institute of Internal Auditors, The Definition of Internal Auditing, 2024. https://www.theiia.org/en/standards/what-are-the-standards/definition-of-internal-audit/.

[57] ISACA, Auditing Artificial Intelligence, 2018. https://www.isaca.org/bookstore/bookstore-wht_papers-digital/whpaai.

[58] J. Jiang, S. Vosoughi, Not judging a user by their cover: understanding harm in multi-modal processing within social media research, in: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, 2020, pp. 6–12.

[59] A. Jobin, I. Marcello, V. Effy, The global landscape of AI ethics guidelines, Nat. Mach. Intell. 1 (9) (2019) 389–399.

[60] M. Jovanovic, M. Campbell, Generative Artificial Intelligence: trends and prospects, Computer. (Long. Beach. Calif) 55 (10) (2022) 107–112, https://doi.org/10.1109/mc.2022.3192720.

[61] A. Kaplan, M. Haenlein, Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence, Bus. Horiz. 62 (1) (2019) 15–25.

[62] M. Katell, B. Herman, C. Binz, M. Young, V. Guetler, D. Raz, D. Dailey, A. Tam, P. M. Krafft, Toward situated interventions for algorithmic equity: lessons from the field, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 45–55.

[63] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy Artificial Intelligence: a review, ACM. Comput. Surv. 55 (2) (2022) 1–38, https://doi.org/10.1145/3491209.

[64] M. Kearns, S. Neel, A. Roth, Z.S. Wu, An Empirical Study of Rich Subgroup Fairness for Machine Learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 100–109.

[65] M. Kim, A. Ghorbani, J. Zou, Multiaccuracy: black-box post-processing for fairness in classification, in: AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 247–254.

[66] B. Knowles, J.T. Richards, The sanction of authority: promoting public trust in AI, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 262–271, https://doi.org/10.1145/3442188.3445890.

[67] A. Koshiyama, E. Kazim, P. Treleaven, P. Rai, L. Szpruch, G. Pavey, G. Ahamat, F. Leutner, R. Goebel, A. Knight, J. Adams, C. Hitrova, J. Barnett, P. Nachev, D. Barber, T. Chamorro-Premuzic, K. Klemmer, M. Gregorovic, S. Khan, E. Lomas, Toward Algorithm Auditing: a Survey on Managing Legal, Ethical, and Technological Risks of AI, ML, and Associated Algorithms (SSRN Scholarly Paper No. ID 3778998), Social Science Research Network, 2021.

[68] KPMG, A Risk and Controls Matrix, 2018. https://pair-code.github.io/what-if-tool/index.html.

[69] J. Kroll, J. Huey, S. Barocas, E. Felten, J. Reidenberg, D. Robinson, H. Yu, Accountable Algorithms, 165, University of Pennsylvania Law Review, 2017, pp. 633–705.

[70] J.A. Kroll, Outlining traceability: a principle for operationalizing accountability in computing systems, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 758–771, https://doi.org/10.1145/3442188.3445937.

[71] O.N. Kulkarni, V. Patil, V.K. Singh, P.K. Atrey, Accuracy and fairness in pupil detection algorithm, in: 2021 IEEE Seventh International Conference on Multimedia Big Data (BigMM, 2021, pp. 17–24, https://doi.org/10.1109/BigMM52142.2021.00011.

[72] J. Kulshrestha, M. Eslami, J. Messias, M. Zafar, S. Ghosh, K. Gummadi, K. Karahalios, Quantifying search bias: investigating sources of bias for political searches in social media, in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 417–432.

[73] K. Kyriakou, P. Barlas, S. Kleanthous, J. Otterbacher, Fairness in proprietary image tagging algorithms: a cross-platform audit on people images, in: Proceedings of the Thirteenth International AAAI Conference on Web and Social Media 13, 2019, pp. 313–322.

[74] S. Laato, M. Mäntymäki, M. Minkkinen, T. Birkstedt, A.K.M.N. Islam, D. Dennehy, Integrating machine learning with software development lifecycles: insights from experts, in: ECIS 2022 Proceedings, ECIS, Timișoara, Romania, 2022.

[75] S. Laato, M. Tiainen, A.K.M. Najmul Islam, M Mäntymäki, How to explain AI systems to end users: a systematic literature review and research agenda, Internet Res. 32 (7) (2022) 1–31, https://doi.org/10.1108/intr-08-2021-0600.

[76] R. LaBrie, G. Steinke, Toward a framework for ethical audits of AI algorithms, in: Twenty-fifth Americas Conference on Information Systems, 2019, pp. 33–44.

[77] R.N. Landers, T.S. Behrend, Auditing the AI auditors: a framework for evaluating fairness and bias in high-stakes AI predictive models, Am. Psycholog. (2022), https://doi.org/10.1037/amp0000972.

[78] S. Lebovitz, N. Levina, H. Lifshitz-Assa, Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what, MIS Q. 45 (3) (2021) 1501–1526, https://doi.org/10.25300/MISQ/2021/16564.

[79] M.C.M. Lee, H. Scheepers, A.K.H. Lui, E.W.T. Ngai, The implementation of Artificial Intelligence in organizations: a systematic literature review, Inform. Manag. (2023), https://doi.org/10.1016/j.im.2023.103816.

[80] J. Li, M. Li, X. Wang, J. Bennett Thatcher, Strategic directions for AI: the role of CIOs and boards of directors, MIS Q. 45 (3) (2021) 1603–1644, https://doi.org/10.25300/MISQ/2021/16523.

[81] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: from principles to practices, ACM. Comput. Surv. 55 (9) (2023) 1–46, https://doi.org/10.1145/3555803.

[82] M. Loi, M. Spielkamp, Toward accountability in the use of artificial intelligence for public administrations, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 757–766, https://doi.org/10.1145/3461702.3462631.

[83] G. Malgieri, G. Comande, Why a right to legibility of automated decision-making exists in the general data protection regulation, Int. Data Priv. Law 7 (4) (2017) 243–265.

[84] M. Martinez, A. Fernandez, AI in recruiting. multi-agent systems architecture for ethical and legal auditing, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019, pp. 6428–6429.

[85] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM. Comput. Surv. 54 (6) (2019) 1–35.

[86] R. Mehrotra, A. Anderson, F. Diaz, A. Sharma, H. Wallach, E. Yilmaz, Auditing search engines for differential satisfaction across demographics, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 626–633.

[87] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable Artificial Intelligence: objectives, stakeholders, and future research opportunities, Inform. Syst. Manag. 39 (1) (2022) 53–63, https://doi.org/10.1080/10580530.2020.1849465.

[88] D. Metaxa, M.A. Gan, S. Goh, J. Hancock, J.A. Landay, An image of society: gender and racial representation and impact in image search results for occupations, Proc. ACM. Hum. Comput. Interact. 5 (CSCW1) (2021) 1–23, https://doi.org/10.1145/3449100.

[89] J. Metcalf, E. Moss, E.A. Watkins, R. Singh, M.C. Elish, Algorithmic impact assessments and accountability: the co-construction of impacts, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2021, pp. 735–746.

[90] P. Mikalef, K. Conboy, J.E. Lundström, A. Popovič, Thinking responsibly about responsible AI and 'the dark side' of AI, Eur. J. Inform. Syst. 31 (3) (2022) 257–268, https://doi.org/10.1080/0960085x.2022.2026621.

[91] P. Mikalef, M. Gupta, Artificial intelligence capability: conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance, Inform. Manag. 58 (3) (2021) 103434, https://doi.org/10.1016/j.im.2021.103434.

[92] M. Minkkinen, M. Mäntymäki, Discerning between the "Easy" and "Hard" problems of AI governance, IEEE Trans. Technol. Soc. 4 (2) (2023) 188–194, https://doi.org/10.1109/TTS.2023.3267382.

[93] M. Minkkinen, A. Niukkanen, M. Mäntymäki, What About investors? ESG Analyses As Tools For Ethics-Based AI Auditing, a, AI & Society, 2022, https://doi.org/10.1007/s00146-022-01415-0.

[94] M. Minkkinen, M.P. Zimmer, M. Mäntymäki, Co-shaping an ecosystem for responsible AI: five types of expectation work in response to a technological frame, Inform. Syst. Front. (2022), https://doi.org/10.1007/s10796-022-10269-2.

[95] B. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: mapping the debate, Big. Data Soc. 3 (2) (2016) 1–21.

[96] B. Mittelstadt, Principles alone cannot guarantee ethical AI, Nat. Mach. Intell. 1 (11) (2019) 501–507.

[97] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Phys. Ther. 89 (9) (2009) 873–880.

[98] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM. Trans. Interact. Intell. Syst. 11 (3–4) (2021) 1–45, https://doi.org/10.1145/3387166.

[99] V.C. Müller, Ethics of artificial intelligence and robotics, in: E. Zalta (Ed.), Stanford Encyclopedia of Philosophy, CSLI, Stanford University, Palo Alto, Cal., 2020.

[100] J. Mökander, M. Axente, Ethics-based Auditing of Automated Decision-Making systems: Intervention points and Policy Implications, AI & Society, 2021, https://doi.org/10.1007/s00146-021-01286-x.

[101] J. Mökander, M. Axente, F. Casolari, L. Floridi, Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI Regulation, Mind. Mach. (Dordr.) 32 (2) (2022) 241–268, https://doi.org/10.1007/s11023-021-09577-4.

[102] J. Mökander, J. Morley, M. Taddeo, L. Floridi, Ethics-based auditing of automated decision-making systems: nature, scope, and limitations, Sci. Eng. Ethics 27 (4) (2021) 44, https://doi.org/10.1007/s11948-021-00319-4.

[103] J. Mökander, L. Floridi, Ethics-based auditing to develop trustworthy AI, Mind. Mach. (Dordr.) 31 (2) (2021) 323–327, https://doi.org/10.1007/s11023-021-09557-8.

[104] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, 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 (2020) 2141–2168.

[105] M. Mäntymäki, M. Minkkinen, T. Birkstedt, M. Viljanen, Defining organizational AI governance, AI. Ethic. (2022), https://doi.org/10.1007/s43681-022-00143-x.

[106] J. Mökander, J. Morley, M. Taddeo, L. Floridi, Ethics-based auditing of automated decision-making systems: nature, scope, and limitations, Sci. Eng. Ethics 27 (44) (2021).

[107] J. Mökander, L. Floridi, Operationalizing AI governance through ethics-based auditing: an industry case study, AI. Ethic. (2022), https://doi.org/10.1007/s43681-022-00171-7.

[108] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Sci. (1979) 366 (6,464) (2019) 447–453.

[109] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, FairLens: auditing black-box clinical decision support systems, Inf. Process. Manage 58 (5) (2021) 102657, https://doi.org/10.1016/j.ipm.2021.102657.

[110] E. Papagiannidis, I.M. Enholm, C. Dremel, P. Mikalef, J. Krogstie, Toward AI Governance: identifying best practices and potential barriers and outcomes, Inform. Syst. Front. 25 (1) (2023) 123–141, https://doi.org/10.1007/s10796-022-10251-y.

[111] O. Papakyriakopoulos, J. Serrano, S. Hegelich, F. Marco, Bias in word embeddings, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 446–457.

[112] O. Papakyriakopoulos, A.M. Mboya, Beyond algorithmic bias: a socio-computational interrogation of the Google Search by image algorithm, Soc. Sci. Comput. Rev. (2022), https://doi.org/10.1177/08944393211073169, 089443932110731.

[113] K.H.S. Pickett, The Internal Auditing Handbook, John Wiley & Sons, 2010.

[114] S. Park, S. Kim, Y. Lim, Fairness audit of machine learning models with confidential computing, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3488–3499, https://doi.org/10.1145/3485447.3512244.

[115] H. Park, D. Ahn, K. Hosanagar, J. Lee, Designing fair AI in human resource management: understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions, in: CHI Conference on Human Factors in Computing Systems 1–22, 2022, https://doi.org/10.1145/3491102.3517672.

[116] J. Quedado, A. Zolyomi, A. Mashhadi, A case study of integrating fairness visualization tools in machine learning education, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts 1–7, 2022, https://doi.org/10.1145/3491101.3503568.

[117] I. Raji, J. Buolamwini, Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products, in: AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

[118] I. Raji, M. Mitchell, J. Buolamwini, J. Lee, T. Gebru, E. Denton, Saving face: investigating the ethical concerns of facial recognition auditing, in: Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 145–151.

[119] I. Raji, M. Mitchell, J. Smith-Loud, A. Smart, T. Gebru, D. Theron, B. Hutchinson, P. Barner, Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 33–44.

[120] B. Rakova, J. Yang, H. Cramer, R. Chowdhury, Where Responsible AI Meets Reality: practitioner Perspectives on Enablers for Shifting Organizational Practices, Proc. ACM. Hum. Comput. Interact. 5 (CSCW1) (2021) 1–23, https://doi.org/10.1145/3449081.

[121] C. Reed, E. Kennedy, S. Silva, Responsibility, autonomy and accountability: legal liability for machine learning, SSRN Electronic Journal (2016).

[122] R. Robertson, S. Jiang, K. Joseph, L. Friedland, D. Lazer, C. Wilson, Auditing partisan audience bias within Google Search, Proc. ACM. Hum. Comput. Interact. 2 (2018) 1–22.

[123] D.M.F. Saldanha, C.N. Dias, S. Guillaumon, Transparency and accountability in digital public services: learning from the Brazilian cases, Gov. Inf. Q. 39 (2) (2022) 101680, https://doi.org/10.1016/j.giq.2022.101680.

[124] C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, Auditing algorithms: research methods for detecting discrimination on internet platforms, in: A preconference at the 64th Annual Meeting of the International Communication Association, 2014.

[125] P. Sapiezynski, W. Zeng, R. Robertson, A. Mislove, C. Wilson, Quantifying the impact of user attention on fair group representation in ranked lists, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 553–562.

[126] M. Scheuerman, K. Wade, C. Lustig, J. Brubaker, How we've taught algorithms to see identity: constructing race and gender in image databases for facial analysis, Proc. ACM. Hum. Comput. Interact. 4 (1) (2020) 1–35.

[127] J. Schneider, R. Abraham, C. Meske, J. Vom Brocke, Artificial intelligence governance for businesses, Inform. Syst. Manag. 40 (3) (2023) 229–249, https://doi.org/10.1080/10580530.2022.2085825.

[128] N. Schöppl, M. Taddeo, L. Floridi, Ethics auditing: lessons from business ethics for ethics auditing of AI, in: The 2021 Yearbook of the Digital Ethics Lab: Digital Ethics Lab Yearbook, Springer International Publishing, 2022, pp. 209–227, https://doi.org/10.1007/978-3-031-09846-8_13.

[129] C. Seidelin, T. Moreau, I. Shklovski, N. Holten Møller, Auditing risk prediction of long-term unemployment, in: Proceedings of the ACM on Human-Computer Interaction 6, 2022, pp. 1–12, https://doi.org/10.1145/3492827.

[130] A. Selcuk, A guide for systematic reviews: PRISMA, Turk. Arch. Otorhinolaryngol. 57 (1) (2019) 57–58.

[131] A. Seppälä, T. Birkstedt, M. Mäntymäki, From ethical AI principles to governed AI, in: Proceedings of the 42nd International Conference on Information Systems (ICIS2021). International Conference on Information Systems (ICIS), Austin, Texas, 2021. https://aisel.aisnet.org/icis2021/ai_business/ai_business/10/.

[132] H. Shen, A. DeVos, M. Eslami, K. Holstein, Everyday algorithm auditing: understanding the power of everyday users in surfacing harmful algorithmic behaviors, Proc. ACM. Hum. Comput. Interact. 5 (CSCW2) (2021) 1–29, https://doi.org/10.1145/3479577.

[133] B. Shneiderman, Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems, ACM. Trans. Interact. Intell. Syst. 10 (4) (2020) 1–31.

[134] A. Shulner-Tal, T. Kuflik, D. Kliger, Fairness, explainability, and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system, Ethic. Inf. Technol. 24 (1) (2022) 2, https://doi.org/10.1007/s10676-022-09623-4.

[135] J. Singh, J. Cobbe, C. Norval, Decision provenance: harnessing data flow for accountable systems, IEEE Access. 7 (2019) 6562–6574.

[136] J. Singh, I. Walden, J. Crowcroft, J. Bacon, Responsibility & machine learning: part of a process, SSRN Electron. J. (2016). http://dx.doi.org/10.2139/ssrn.2860048.

[137] V. Singh, C. Hofenbitzer, Fairness across network positions in cyberbullying detection algorithms, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 557–559.

[138] M. Sloane, E. Moss, R. Chowdhury, A Silicon Valley love triangle: hiring algorithms, pseudo-science, and the quest for auditability, Patterns 3 (2) (2022) 100425, https://doi.org/10.1016/j.patter.2021.100425.

[139] B.C. Stahl, Artificial Intelligence For a Better future: An ecosystem Perspective On the Ethics of AI and Emerging Digital Technologies (SpringerBriefs in Research and Innovation Governance), Springer International Publishing, 2021.

[140] T. Sturm, J. Gerlacha, L. Pumplun, N. Mesbah, F. Peters, C. Tauchert, N. Nan, P. Buxmann, Coordinating human and machine learning for effective organization learning, MIS Quarterly 45 (3) (2021) 1581–1602, https://doi.org/10.25300/MISQ/2021/16543.

[141] I. Sulaimon, A. Ghoneim, M. Alrashoud, A new reinforcement learning-based framework for unbiased autonomous software systems, in: 8th International Conference on Modeling Simulation and Applied Optimization, 2019, pp. 1–6.

[142] Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2017). Detecting Bias in Black-Box Models Using Transparent Model Distillation. arXiv preprint arXiv:1710.06169..

[143] M. Teodorescu, L. Morse, Y. Awwad, G. Kane, Failures of fairness in automation require a deeper understanding of human-ML augmentation, MIS Quarterly 45 (3) (2021) 1483–1500, https://doi.org/10.25300/MISQ/2021/16535.

[144] S. Toapanta, J. Monar, L. Gallegos, Prototype to perform audit in social networks to determine cyberbullying, in: *2020 Fourth World Conference on Smart Trends in Systems, Security, and Sustainability*, 2020, pp. 145–153.

[145] N. Tsakalakis, S. Stalla-Bourdillon, L. Carmichael, T.D. Huynh, L. Moreau, A. Helal, The dual function of explanations: why it is useful to compute explanations, Comput. Law Secur. Rev. 41 (2021) 105527, https://doi.org/10.1016/j.clsr.2020.105527.

[146] B. Vecchione, K. Levy, S. Barocas, Algorithmic auditing and social justice: lessons from the history of audit studies, Equ. Access Algorith. Mech. Optim. (2021) 1–9, https://doi.org/10.1145/3465416.3483294.

[147] K. Werder, B. Ramesh, R. Zhang, Establishing data provenance for responsible Artificial Intelligence systems, ACM. Trans. Manag. Inf. Syst. 13 (2) (2022) 1–23, https://doi.org/10.1145/3503488.

[148] M. Wieringa, What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 1–18.

[149] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, Building and auditing fair algorithms: a case study in candidate screening, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 666–677, https://doi.org/10.1145/3442188.3445928.

[150] H. Zhang, N. Shahbazi, X. Chu, A. Asudeh, FairRover: explorative model building for fair and responsible machine learning, in: Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning, 2021, pp. 1–10, https://doi.org/10.1145/3462462.3468882.

[151] J. Zhang, C. Wan, C. Zhang, X. Guo, T. Lu, Auditing images collected by sensors in ambient intelligence systems with privacy and high efficiency, J. Supercomput. 77 (11) (2021) 12771–12789, https://doi.org/10.1007/s11227-021-03738-z.

Joakim Laine is a doctoral candidate in Information Systems at Turku School of Economics. He holds an M.*Sc*. (Econ. & Bus. Adm.) from University of Turku. His-research focuses on the auditing and governance of artificial intelligence.

Matti Minkkinen is a Senior Researcher in Information Systems at Turku School of Economics. He holds a Ph.D. in Futures Studies, and his research interests cover the dynamics of expectations and technologies and the impacts of digital transformation on individuals and organizations. He has several years of research and teaching experience on the interplay between future visions and socio-technical change as well as foresight methods. Minkkinen's recent research covers responsible artificial intelligence as a socio-technical and networked phenomenon, European debates on privacy protection, systemic foresight processes, and futures consciousness as a human capacity. His-research has been published in journals such as *Technological Forecasting & Social Change, New Media & Society, Information Systems Frontiers, and Personality and Individual Differences.*

Matti Mäntymäki is a Professor of Information Systems Science at University of Turku, Finland. His-research interests cover a broad range of psychosocial, organizational, and business implications of digitalization, including governance and social responsibility of artificial intelligence. He has authored more than 100 peer-reviewed papers, published in outlets such as *Information Systems Journal, Technological Forecasting & Social Change, International Journal of Information Management, Journal of Business Research, Information Systems Frontiers, Information Technology & People, Computers in Human Behavior, Journal of Systems & Software,* and *Communications of the Association for Information Systems*, among others.