UNIVERSITY OF TURKU
Turku School of Economics

# Ethics-based AI auditing core drivers and dimensions: A systematic literature review

Information Systems Science
Master's thesis

Author:
Joakim Laine

Supervisors:
Ph.D. Matti Minkkinen
D.Sc. Samuli Laato

29.10.2021
Turku

Master's thesis

**Subject**: Information Systems Science
**Authors**: Joakim Laine
**Title**: Ethics-based AI auditing core drivers and dimensions: A systematic literature review
**Supervisor(s)**: Ph.D. Matti Minkkinen, D.Sc Samuli Laato
**Number of pages**: 82 pages
**Date**: 29.10.2021

This thesis provides a systematic literature review (SLR) of ethics-based AI auditing research. The review's main goals are to report the current status of AI auditing academic literature and provide findings addressing the review objectives. The review incorporated 50 articles presenting ethics-based AI auditing. The SLR findings indicate that the AI auditing field is still new and rising. Most of the studies were conference proceeding published either 2019 or 2020. Therefore, there was a demand for a SLR work as the AI auditing field was wide and unorganized.

Based on the SLR findings, fairness, transparency, non-maleficence and responsibility are the most important principles for the ethics-based AI auditing. Other commonly identified principles were privacy, beneficence, freedom and autonomy and trust. These principles were interpreted to belong to either drivers or dimensions depending on whether something is audited directly or whether achieving ethics is a desired outcome.

The findings also suggest that the external AI auditing leads the ethics-based AI auditing discussion. Majority of the papers dealt specifically with external AI auditing. The most important stakeholders were recognized to be researchers, developers and deployers, regulators, auditors, users and individuals and companies. Roles of the stakeholders varied depending on whether they are proposed to conduct AI audits or whether they are in the position of beneficiary.

**Key words**: artificial intelligence, AI, auditing, ethics, machine learning, principles, systematic review, SLR.

Pro gradu -tutkielma

**Oppiaine**: Tietojärjestelmätiede
**Tekijät**: Joakim Laine
**Otsikko**: Ethics-based AI auditing core drivers and dimensions: A systematic literature review
**Ohjaajat**: FT Matti Minkkinen, TkT Samuli Laato
**Sivumäärä**: 82 sivua
**Päivämäärä**: 29.10.2021

Tässä Pro gradu -tutkielmassa esitellään systemaattinen kirjallisuuskatsaus etiikkalähtöiseen tekoälyn auditointiin. Kirjallisuuskatsauksen keskeisimmät tavoitteet ovat esittää tämänhetkinen tila tekoälyn auditoinnin akateemisesta kirjallisuudesta sekä esittää keskeisimmät löydökset tutkielman tavoitteiden mukaisesti. Kirjallisuuskatsaus sisälsi 50 artikkelia, mitkä käsittelivät etiikkalähtöistä tekoälyn auditointia. Systemaattisen kirjallisuuskatsauksen löydökset osoittivat, että tekoälyn auditoinnin ala on edelleen uusi ja kasvava. Suurin osa julkaisuista oli konferenssipapereita vuosilta 2019-2020. Ala on myös laaja sekä epäorganisoitu, joten systemaattiselle kirjallisuuskatsaukselle oli kysyntää.

Löydöksien perusteella reiluus, läpinäkyvyys, ei-haitallisuus sekä vastuullisuus ovat tärkeimmät periaatteet etiikkalähtöiseen tekoälyn auditointiin. Muut yleisesti tunnistetut periaatteet olivat yksityisyys, hyvyys, vapaus ja autonomia sekä luottamus. Nämä periaatteet tulkittiin kuuluvaksi joko ajureihin tai dimensioihin sen perusteella auditoitiinko periaatetta suoraan vai oliko periaatteen saavuttaminen auditoinnin toivottu tulos.

Löydökset osoittivat myös, että ulkoinen auditointi hallitsee tämänhetkistä keskustelua etiikkalähtöisessä tekoälyn auditoinnissa. Valtaosa papereista käsitteli erityisesti ulkoista tekoälyn auditointia. Lisäksi tärkeimmät sidosryhmät tunnistettiin. Nämä olivat tutkijat, järjestelmän kehittäjät, lainvalvojat, auditoijat, käyttäjät sekä yksilöt ja organisaatiot. Heidän roolinsa vaihtelivat sen perusteella vastasivatko he tekoälyn auditoinnin toteuttamisesta vai kuuluivatko he tekoälyn auditoinnin edunsaajiin.

**Avainsanat:** tekoäly, AI, auditointi, etiikka, koneoppiminen, periaatteet, systemaattinen kirjallisuuskatsaus, SLR.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# 1 Introduction

## 1.1 Research background and motivation

Artificial intelligence (AI) is a rapidly growing field with an increasing growth in the capabilities and applications. This means that more and more companies are looking into AI, and it opens opportunities for new applications and services, and for enhancing existing systems (Dignum 2019). However, this has brought many new challenges for AI, which is why the importance of AI auditing is rising. We need to understand what AI is and what it is not, but more importantly, we need to understand what it can do, how we can ensure a beneficial use of AI and how we put in place the social and technical constructs that ensure responsibilities and trust for the AI systems. (Dignum 2019.) Ethical, technical, social and legal layers include, for example, holding algorithm and data accountable to standards, establishing ethical principles and acceptable practices and legal requirements (LaBrie & Steinke 2019). All these challenges complicate practicing business in the field of AI. The main focus of this thesis is on ethics-based AI auditing. Brown et al. (2020) define it as an "*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 give rise to these negative impacts.*"

AI governance and auditing are key tools for answering the challenges mentioned above. Butcher & Beridze (2019) state that AI governance can be characterized as a variety of solutions, tools, and levers that influence AI development and applications. This may include, for example, promoting norms, ethics and values frameworks. In this thesis I will focus on AI auditing which partially overlaps AI governance. Marques and Santos (2017) mention that The Institute of Internal Auditors defined auditing as an evaluation of effectiveness of control, risk management and governance processes which are designed to add value and improve operations of organizations while achieving their objectives. The algorithmic auditing literature has examined, for example, search engines, online maps, social networks, e-commerce, online advertising and online job boards (Chen et al. 2018). It is worth noting that neither AI governance nor AI auditing are established concepts. According to Brown et al. (2020), auditors collect and analyze data about the behavior of an algorithm and then uses the data to find out whether people are negatively impacted by the behavior of the algorithm. Problems with transparency and explainability are typical characteristics of machine learning (ML) models, which is why it is so important to understand and utilize the worth of AI auditing (Kroll et al. 2016).

Decisions made by AI have raised a lot of criticism because of their potential discrimination and opacity, which is why regulations, drivers and dimensions need to be addressed. Therefore, it is important to look at which factors are guiding the ethics-based AI auditing and how to evaluate them. If an algorithm gives people some sort of scoring system, auditing mostly focuses on issues like unfair treatment and bias or if an algorithm tracks online behavior the focus is primarily on transparency or autonomy issues. (Brown 2020.) Barlas et al. (2019), for example, give an example about algorithm discrimination when in 2015 a black software engineer was labeled as a gorilla by Google photos, and in their other study Barlas et al. (2019) are writing about an incident in 2017 when Apple had to give refunds for Chinese users because FaceID technology could not distinguish between Asian faces.

Some of these concerns are addressed via regulations. Europe in particular has been prominent here. The most prominent example has been the EU General Data Protection Regulation (GDPR). It gives users the right to know how their data is processed (Song & Shmatikov 2019) and obligates data controllers to be able to demonstrate compliance with its various requirements (Sing et al. 2018). More specifically, GDPR article 22, for example, defines the regulatory framework for automated individual-level decision-making which drives companies and organizations to audit their algorithmic services, technologies and procedures (Clavell et al. 2020). In addition, recommendations like the European Commission's Ethics Guidelines for Trustworthy AI guides organizations towards making automated decisions explainable and transparent (Domingo-Ferrer et al. 2019). New regulations are also rising. On April 21, the European Commission unveiled the first-ever legal framework on AI called the Artificial Intelligence Act, which addresses risks from the use of AI and promotes innovation (Gaumond 2021). It is still just a proposal, but it reflects the direction of the development. There are also actors independent of the European Commission which guide the development of AI. One of the most well-known is the High-level Expert Group on Artificial Intelligence. It is a group of 52 experts appointed by the European Commission with a mission to provide advice on AI strategies. (AI HLEG 2019.)

Carrier & Brown (2021) give two kind of term definitions of AI auditing. Casually stated audit is just an in-depth examination about fairness, accountability and transparency of an algorithm. However, on a more professional level it is understood as a robust, long-established, set of principles. They are arguing that the AI ethics industry has created wrongful terminology causing harmful confusion for the public and for the owners of algorithms. For example, Raji et al. (2020) state that audit is defined as "*an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures*" by IEEE Standard for software development.

Li et al. (2019) list main ethical issues and causes of artificial intelligence. There are human rights ethics, moral ethics, prejudice ethics, information ethics, liability ethics and environmental ethics. These link highly on key elements of ethics-based AI auditing like fairness, bias, regulations, privacy, utilization and responsibility. On technical development side of AI auditing, Domanski (2019) brings up five core principles. He argues that responsibility, explainability, accuracy, auditability and fairness should be included in the technical auditing. Along the same lines is also China, which launched eight fundamental principles for governance of responsible artificial intelligence. Those principles are Fairness and Justice, Harmony and Human-friendly, Respect for Privacy, Inclusion and Sharing, Safety and Controllability, Shared Responsibility, Open and Collaboration, Agile Governance. (Zhang & Gao, 2019.) In addition to these, Jobin et al. (2019) listed five core ethical principles around ethics-based AI which were transparency, justice and fairness, non-maleficence, responsibility and privacy. There are clear similarities from all these lists.

AI governance literature is a relatively wide but an unorganized area (Butcher & Beridze 2019) which can be said also about AI auditing literature as they are relatively close concepts. Different frameworks have been generated, for example, PAPA framework (LaBrie & Steinke 2019), TuringBox framework (Epstein et al. 2018) or SMACTR framework (Raji et al. 2020), but the literature has insufficiently mapped out basic questions. Therefore, this thesis aims to conceptualize ethics-based AI auditing academic literature, recognize drivers and dimensions leading ethics-based AI auditing and to identify actors discussed in ethics-based AI auditing literature and clarify their roles. Drivers describe questions what and why about concepts which are driving the field of AI auditing. Dimensions on the other hand describe how question about measurements which are leading AI auditing in a certain direction. These drivers and dimensions are linked with identified ethical principles of ethics-based AI auditing.

There are multiple reasons why people conduct AI auditing. According to Kazim (2021), there exists a bifurcated approach for AI auditing. On the one hand, there is consultancy, where companies seek guidance, an ethics strategy or reputational boost. On the other hand, there is a forensic audit, where an auditor will investigate a company's data and algorithms. Carrier & Brown (2021) published a taxonomy of AI audit, assurance and assessment. They write about confusion and uncertainty what AI audit is and how industries misuse the term. About the approaches of AI auditing, they write “"*Audit is a form of Assurance that uses Rules and Laws. Assurance is a slightly softer version of the same service using rules, guidelines, and standards that have slightly less objectivity and often are not codified in law. Audit is a specialized subset of Assurance. Assurance does not necessarily mean 'audit*." Companies seek high-level guidance, an ethics strategy or reputational boost. There are also worries that AI auditing is used for legitimizing harmful technologies or whitewashing companies' reputations. The AI auditing industry would benefit from formal and standard principles of AI ethics but as different actors from society all see and understand problems differently, it is a great challenge for this industry. (Clarke 2021.)

*“We don't even know what ‘bias’ means or what ‘harm’ means, so that is a real concern.” Mona Sloane, Senior researcher NYU Centre for Responsible AI* (Clarke 2021).

To achieve the most comprehensive understanding of ethics-based AI auditing, I conducted a systematic literature research (SLR). Literature review showed key themes around AI auditing, but an accurate overall definition was difficult to construct. This thesis tries to construct a more structured description of ethics-based AI auditing focusing on research gaps presented earlier. Batarseh et al. (2021), for example, made a systematic review on AI assurance aiming to provide a structured alternative to the landscape. They managed to develop a new definition, contrast and tabulate new methods and evaluate and compare existing methods with new metrics system. Existing definitions of assurance showed that two main AI components, the data and the algorithm, are the main pillars of AI assurance which led to their definition: “*A process that is applied at all stages of the*

*AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users*." This shows that systematic review is an efficient way to make a structured and objective description of the chosen topic.

Technology companies have joined the discussion around ethical AI and AI auditing. Google, for example, published What-if Tool which aimed to increase understanding of ML systems performance across a wide range of inputs (Google 2020). Also, Microsoft published their toolkit assessing and improving fairness in AI called Fairlearn (Bird et al. 2020). Defining principles regarding ethics-based AI auditing and self-regulation are topics which have increased interest around them. However, there is little consensus what ethical means in the AI context or what audit means in that context (Clarke 2021). Businesses are interested in moral implications of algorithms and other ethical pitfalls. The need of meeting ethical requirements is rising in ethics-based AI auditing field, but universal AI ethics frameworks are still missing.  Negative biasing around AI algorithmic systems, image analysis technology influences, industry standards and regulations around AI brings challenges and opportunities for the companies and for the society (Mittelstadt 2019). This was also the motivation for this SLR and directed the research questions and objectives.

This thesis will focus on these issues. First, I will show a brief definition of traditional auditing and ethical auditing. In chapter two I will present the methodology of the thesis and how the SLR was executed. Chapter three will present the results. Section 3.1 shows descriptive data of the sample, section 3.2 presents findings of the ethics-based drivers and dimensions, and section 3.3 presents the actor-based approach where the focus is on stakeholders of ethics-based AI auditing and actors who are conducting ethics-based AI auditing. Finally, chapters 5 and 6 present the discussion and conclusions.

## 1.2 Research questions

The goal of this thesis is to answer the main research question via a systematic literature review:

1) How are ethics-based AI auditing principles discussed in the academic AI auditing literature?

With the support of the additional research questions:

2) What are the most important drivers and dimensions of ethics-based AI auditing?

3) Who are the stakeholders of AI auditing?

4) Who are proposed to conduct the AI audits?

## 1.3 Definitions of auditing

### 1.3.1 Traditional IT auditing

In order to understand AI auditing, we need to understand traditional IT auditing. Historically, auditors were mostly accountant employees to give opinion of a company's accounts and the scope of audits were mostly on finance, for example taxes, accounting processes and accounting systems. AI auditing is a new and unsettled field, but traditional auditing procedures have been relied on for many years. Importance of auditing grows alongside with growth of IT. Nowadays, the same effect is happening between AI auditing and the growth of AI. IT auditing focus on that the program is doing what it is supposed to do, considering that today's issues and tomorrow's threats are taken into account. (Hinson 2007.) The IT field is moving constantly forward and new technical advances and information capabilities are rising.

Magee (2021) defined IT audits as any audit that encompasses both the review and the evaluation of computerized information processing systems, their relation to automated processes and the interfaces among them. Audits are designed to add value and improve operations in organizations to help organizations accomplish its objectives. They can, for example, include assignments that provide assurance or advice with a systematic and disciplined approach. It typically evaluates reports upon the procedures and control environment around the IT systems aiming to improve the effectiveness of risk management, control, and governance processes. More effective management processes can then be reached as audits expose risk entities. (Deloitte 2020.)

Ben Cole (2014) also defined IT auditing as an examination and evaluation of an organization's information technology infrastructure, policies and operations. Main reason for IT auditing is to ensure that information-related controls and processes are working properly. According to him, core objectives for IT auditing are:

- Secure company data by evaluating the systems and processes
- Determine risks to a company's information assets, and help identify methods to minimize those risks.
- Ensure that information management processes follow IT-specific laws, policies and standards.
- Determine inefficiencies in IT systems and associated management.

The primary role with IT auditing was to give assurance for stakeholders and authorities. However, even though this is still partly true with external audits, internal auditing has broader remits. (Hinson 2007.) Eulerich and Kalinichenko (2018) note that external auditors provide assurance regarding quality and adequacy while internal auditors provide assurance regarding operations and risks. Both internal audit departments and external audit firms need to adjust their operations and develop new audit techniques to keep pace with the changing environment. Other forms of auditing have risen, including compliance against legal and regulatory obligations, health and safety policies, environmental protection, quality assurance and management consultancy. While organizations are more and more dependent on IT systems, level of information security threats and vulnerabilities are increasing. Therefore, audit plans involve IT systems and auditors cannot overlook the computer systems and data networks backing the business processes under review. (Hinson 2007.)

### 1.3.2 Ethical auditing

Incorporating ethics into auditing brings new challenges for auditing. No universal definition for ethical audit has been made, but different codes of conduct, laws and regulations and other internal and external controls have been developed for companies, and these are monitored through audits. Ethics auditing is defined by Virovere & Rihma (2008) as an opportunity and agreement to devise a system to inform on ethical corporate behavior. The goal is to increase transparency and credibility of a company’s commitment to ethics. Conflicts in organizations are often caused by violation of ethical principles so ethics auditing helps lower the number of conflicts, and at the same time it allows introducing the moral dimensions in a company's actions. These key dimensions also increase company's trust capital with different stakeholders.

Rosthorn (2000) in turn defines ethical auditing as a regular, complete, and documented measurement of compliance with a company's published policies and procedures. In this regard, Mackenzie (1998) argues that with ethical auditing organizations can contribute with stakeholders to increase their ability to live well. Ethical problems and opportunities are not well understood in corporations and organizations might even think that ethical knowledge is not possible. Ethic is considered as an individual opinion without a real content. However, this approach is also important since it is useful for companies to know stakeholders' opinions even if it is just a matter of opinions. Another discussion is that ethical knowledge is more of a topic for religion or metaphysical speculation than rather than empirical methods like ethical auditing. (Mackenzie 1998.)

Mackenzie (1998) continues that ethical auditing is not for creating ethical knowledge, but rather to discover whether companies are currently complying with the prescriptions of ethical theories. The audit process must collect information about stakeholders, consider the creation of theories and develop methods to test these theories. It has been pursued under headings as political economy, business ethics, economies, business law, accountancy, management theory, and industrial sociology and it can bring an empirical perspective to the ethical understanding and shape the conceptions of ethics found in those fields. Virovere & Rihma (2008) list seven ways how ethical auditing can help organizations to look at their activities and add clarity to its value systems:

1. It clarifies actual values where organizations operate.
2. It helps to measure future improvements by providing a baseline.
3. It can support organizations to meet societal expectations which are not currently being met.
4. Stakeholders get the opportunity to clarify their expectations of company's behavior.
5. Companies can identify specific problem areas.
6. Companies can identify general areas of vulnerabilities, particularly related to lack of openness.
7. It can help organizations to learn about issues which motivates employees.

Opportunities and challenges of ethical auditing keep on rising. External stakeholders and authorities want to meet accountability and transparency expectations and internally it helps to meet ethical objectives of organizations. While AI provides many opportunities, it also brings many challenges for these due to its self-learning nature and problems of determine responsibilities (Leyer & Schneider 2021). Ethical principles, codes of conduct and ethical theories still exist when auditing AI, but drivers and dimensions appear different compared to traditional IT auditing, for example increased autonomy of AI, learning capabilities or predictability, which is why it needs more research.

## 1.4 Theoretical framework

In this study I will focus on AI auditing that aims to ensure that AI is ethical when evaluated against established ethics principles. Alternative directions for AI auditing could, for example, be ensuring legality or efficiency. The division into ethics-based and other category papers was based on AI High-Level Expert Group on Artificial Intelligence (AI HLEG) Ethics guidelines for Trustworthy AI. AI HLEG is an independent group set up by the European commission. Their framework has reached popularity among AI industry; therefore, it works well for a baseline of this division. They divided trustworthy AI into three categories: lawful AI, ethical AI and robust AI. Furthermore, they presented four ethical principles and seven requirements for them. The four principles were respect for human autonomy, prevention of harm, fairness and explicability. (AI HLEG 2019.) Papers whose core focus were some of these principles were marked as ethics-based papers and rest were other category papers. To clarify, other category paper does not mean that the paper itself is non-ethical, but the core focus is somewhere else, for example, in technical aspects.

AI HLEG (2019) is institutionally credible source as they have EU connection, their framework is highly cited, and they are widely known. For these reasons it was chosen to be the baseline for the paper division. For the data analysis, I used Jobin et al. (2019) framework about guidelines on ethical AI. It is academic summary on academic and grey literature; therefore, it is more suitable for the data analysis in this thesis. By combining these two frameworks, I managed to utilize both academic and non-academic literature.

Jobin et al. (2019) have made a highly cited study of the global landscape of AI ethics guidelines. Their study revealed that five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy) are driving the current discussion of ethical AI. In addition, six other ethical principles were identified. These existing principles were taken for the baseline for this work to investigate how these principles are seen in the ethics-based AI auditing literature, and how those principles are pertaining to different actors, why are they important and what issues they might cause.

Jobin et al. (2019) studied what constitutes ethical AI, and what are the ethical requirements, technical standards and best practices for ethical AI. They analyzed current corpus of principles and guidelines of AI ethics revealing how certain principles came up more than others, how those principles linked together, why those are important, what actors they pertain to and how they should be implemented. The research was conducted as a scoping review with the same PRISMA-method as this study but included also grey literature. Used keywords were: AI principles, artificial intelligence principles, AI guidelines, artificial intelligence guidelines, ethical AI and ethical artificial intelligence, and the final sample for the content analysis included 84 documents. As the themes of the study and the sample collection technique corresponded so strongly with the scope of this SLR, the ethical principles identified in that study were taken for the groundwork for analyzing key principles for ethics-based AI auditing.

# 2 Methodology

## 2.1 Research design

In this thesis, I conducted a systematic literature review of AI auditing. Systematic review is defined as: "*a scientific process governed by a set of explicit and demanding rules oriented towards demonstrating comprehensiveness, immunity from bias, and transparency and accountability of technique and execution*" (Davies et al. 2013). It gives a rigorous review of research results (Iden & Eikebrokk 2013), and it enables researchers to perform a systematic, transparent and reproducible synthesis of prior literature (Tandon et al. 2020). AI auditing is not much systematically mapped, so the goal was to systematically scan the AI auditing field. Literature of AI ethics has produced quickly, and it requires harmonization and aggregation. SLR was chosen to be the best tool for scanning the fundamentals of the field and to fulfill the harmonization and aggregation requirements.

This report strategy follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines and PRISMA 2009 checklist (Moher et al. 2009). The idea of a systematic literature review is to systematically search for literature on a specific topic and to synthesize it in order to answer the research questions. Several inclusion and exclusion criteria determine which articles are included in the review, resulting in a sample of articles as comprehensive as possible on a given topic. These articles are then systematically analyzed to synthesize findings and the knowledge of the academic literature and to identify research gaps as well as future research agenda.

The PRISMA framework was created to avoid issues like: (1) the reporting of the review is incomprehensive, (2) method details are not detailed enough, (3) the results contain significant author bias, (4) quality differences are not considered between studies and (5) results are misinterpreted or inadvertent bias. (Selcuk 2019.)

In the current study, the SLR was conducted in three phases. In the first phase I developed a review plan. Research objectives were defined, inclusion and exclusion criteria selected, digital databases were explored, and review process postulated and assessed. In the second phase, four databases were used as sources for research items: Scopus, Web of Science Core Collection, IEE Electronic Library and ACM – Association for Computing Machinery. The search was executed using selected keywords and criteria which brought the original sample. Additional monitoring then shaped the sample into its final form. Last, the data was organized in a suitable form which enabled to synthesize and discuss the findings. Figure 1 demonstrates the search process:

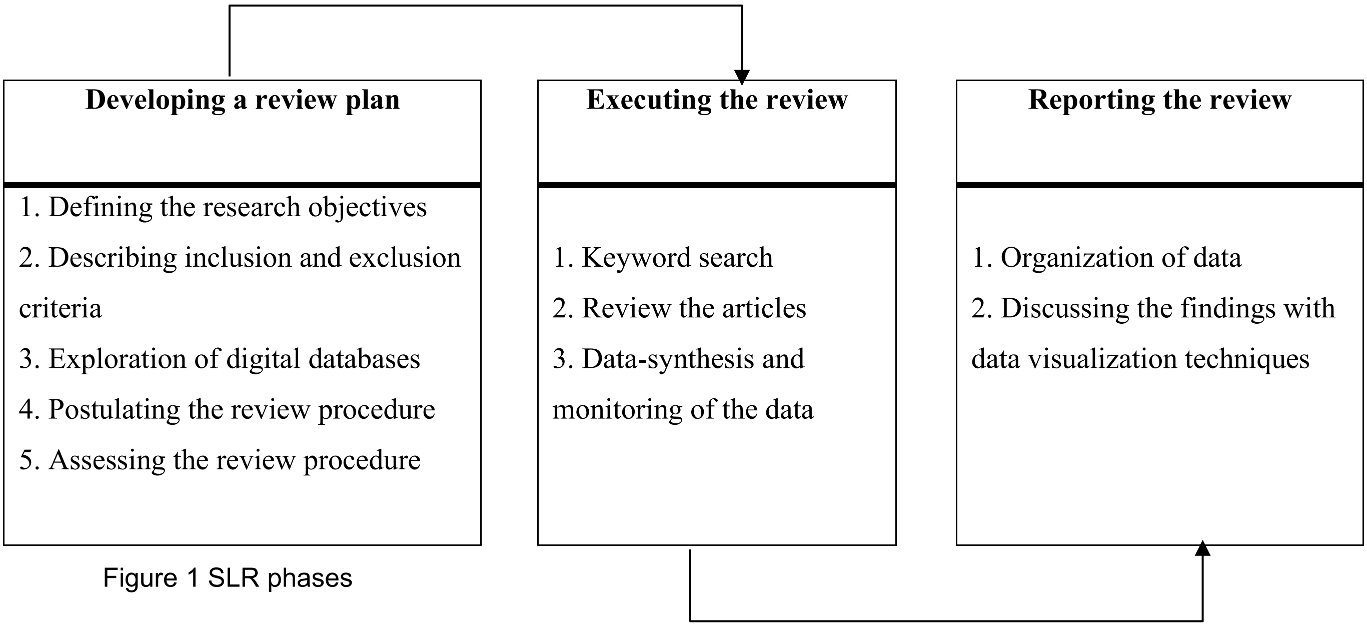

Developing a review plan
1. Defining the research objectives
2. Describing inclusion and exclusion criteria
3. Exploration of digital databases
4. Postulating the review procedure
5. Assessing the review procedure
Executing the review
1. Keyword search
2. Review the articles
3. Data-synthesis and monitoring of the data
Reporting the review
1. Organization of data
2. Discussing the findings with data visualization techniques

Figure 1 SLR phases

Following sections describe these steps in more detail. Section 2.2 presents the full search process. Developing a review plan deals with processes before conducting the search. It focuses on questions why and what, defining the goals and reasons for this SLR and grounds for choosing certain databases and criteria. Executing the review shows the search syntax of keyword searches and specifies chosen keywords. Section 2.3 then shows how the data were handled and what information was found to answer the research objectives mentioned earlier.

With systematically mapping the research field and scanning the fundamentals of AI auditing, I aimed to synthesize the literature, present the ethics-based AI auditing principles and survey how those principles are discussed in the ethics-based AI auditing literature like mentioned in section 1.2. In addition, I intended to look at key drivers and dimensions of ethics-based AI auditing based on the literature which was the second research question. Actor perspective is also an interesting objective, so one goal was to search who are the stakeholders of AI auditing and who are proposed to conduct AI audits.

One important thing to highlight is that there is also literature about how AI is utilized in auditing, in addition to the auditing of AI. For example, Omoteso (2012) studied how to use the application of AI in auditing. However, the focus on this thesis is specifically auditing of AI. This was also addressed in inclusion and exclusion criteria. In this regard, the objective was to review how different auditing actors and audit targets are seen by the literature.

Table 1 Research objectives

| | |
|---|---|
| **#1** | Conceptualize the literature of ethics-based AI auditing |
| **#2** | Key ethical principles of ethics-based AI auditing literature and how those principles are discussed in the literature |
| **#3** | Key drivers and dimensions of ethics-based AI auditing |
| **#4** | Who are the stakeholders of AI auditing |
| **#5** | Who are proposed to conduct AI audits |

## 2.2 Data collection

This section will describe the search process by presenting the number of articles screened for each database and different phases which resulted in the final sample of 58 from the databases and 30 from the backward citation chaining. First, I will go through step by step how I ended up with that sample and then I will present a flowchart which sums up the search process.

Selecting the databases for the search is a critical step. As mentioned, the four databases screened were Scopus, Web of Science Core Collection, IEEE Xplore and ACM digital library. No single database is likely to contain all relevant references which is why I ended up with these four relevant databases. Scopus, for example, indexes IEEE Xplore and ACM. There are dozens of different academic research databases but for this given topic, these four databases seemed most relevant and adding more databases would not have brought significant added value. ACM and IEEE Xplore are specialized in computer science. IEEE Xplore, for example, has over four million records focusing specifically on engineering. Scopus and Web of Science Core Collection are multi-indexed subscription databases and ACM is a digital library with millions of journal and conference papers focusing on computing, so this combination was found sufficient for this SLR.

With Scopus and Web of Science Core Collection, the search covered only titles, abstracts, and keywords. This is because there was no full text search option in these databases. However, IEEE Xplore and ACM digital library covered also full texts. Usually only titles, abstracts and keywords are reviewed in the PRISMA framework but this way I ensured that the auditing field was fully covered, and no relevant articles would be left out even though it required a lot of manual work with false positives. Full text searches result mostly in a sample of articles including only one or two mentions of the given topic, but it is still necessary to do to ensure the most comprehensive outcome.

The databases were reviewed in the following order: 1) Scopus, 2) Web of Science Core Collection, 3) IEEE Xplore, 4) ACM Conference and ACM Journal. ACM Digital Library conference proceedings and journal publications had to be screened separately, as the database search engine does not allow searching for both simultaneously. At the first phase all the search results were downloaded into Excel. After that, duplicates were removed in the same order as the databases were reviewed. Inclusion and exclusion criteria were included into searches but search engines do not always note every criterion for some reason. Therefore, in second phase I removed duplicates, papers not matching inclusion and exclusion criteria, and screened papers based on titles, abstracts and keywords. Last, a full text review was performed. If there were any papers that were not supposed to be in the sample according to the inclusion and exclusion criteria, they were removed in this phase. This resulted in the final sample from databases with a total of 58 articles.

Table 2 presents the inclusion and exclusion criteria. The inclusion criteria in literature search dictated that publications had to be in English and published in a peer reviewed journal or conference proceedings. Books, book chapters, reviews etc. and papers in other languages than English were excluded. Initial search recognized all the audit studies, but third inclusion criteria was taken into account when papers were screened based on titles, abstracts and keywords.

Table 2 Inclusion & Exclusion criteria

| **Inclusion criteria (IC)** | | **Exclusive criteria (EC)** | |
|---|---|---|---|
| **IC#1** | Articles or conference papers only | **EC#1** | Books, book chapters, reviews etc. |
| **IC#2** | Studies published in the English language | **EC#2** | Studies other language than English |
| **IC#3** | Studies address auditing of artificial intelligence | **EC#3** | Focus on something else than auditing of artificial intelligence, e.g. use of artificial intelligence in auditing |
| **IC#4** | Studies published before 2021 | | |

Database screening started with Scopus. First search resulted in a total of 449 articles in Scopus. Like stated earlier, with Scopus the search included only title, abstract and keywords. Search strings had to include the term “auditing” and either “artificial intelligence”, “AI”, “deep learning”, “machine learning”, “black box” or “algorithm”. With the same search, Web of Science Core Collection resulted in a total of 127 articles. IEEE Xplore and ACM Digital Library search included full texts. IEEE resulted in 521 articles, ACM conference 1223 articles and ACM Journal 213 articles. These searches included EC1 and EC2 which means that papers written in other language than English and books, book chapters, reviews etc. were excluded. The total number of articles at this phase was 2533.

The second phase was screening based on titles, abstracts and keywords. IC3 was taken into account which means that papers had to address auditing of AI. Duplicate studies were removed in the order Scopus, Web of Science Core Collection, IEEE Xplore, ACM conference proceedings and ACM Journal publications. These steps considered the second phase resulted in a total of 259 articles. 65 out of these 259 were from Scopus, 7 from Web of Science, 58 from IEEE, 105 from ACM Conference proceedings and 24 from ACM Journal publications. Therefore, 2274 articles were excluded in this phase.

The third phase was screening based on full text. Basically, the 259 articles were analyzed which resulted in the final sample. This phase also confirmed if there were any papers which should not be there as sometimes databases do not apply exclusion criteria correctly or sometimes papers have abstracts in English but full texts in other language. This resulted in the final sample from the databases, a total of 58 articles. 22 of these studies were from Scopus, 3 from Web of Science Core Collection, 6 from IEEE Xplore, 23 from ACM Conference proceedings and 4 from ACM Journal publications. Exact numbers of each phase are presented in table 3:

Table 3 Database number of articles

| Data Source | Phase 1 | Phase 2 | Phase 3 |
|---|---|---|---|
| **Scopus** | 449 | 65 | 22 |
| **Web of Science Core collection** | 127 | 7 | 3 |
| **IEEE Xplore** | 521 | 58 | 6 |
| **ACM Digital Library conference proceedings** | 1223 | 105 | 23 |
| **ACM Digital Library journal publications** | 213 | 24 | 4 |
| **Total** | 2533 | 259 | 58 |

As we can see, ACM Conference proceeding was the most common database, which can be explained with the full text search, following with Scopus. Web of Science, IEEE and ACM Journal publications completed the sample with a few articles each. 201 articles were excluded when moving from reading titles and abstract to the full text analysis. Further analysis was then made for the final sample.

With the selected keywords, I tried to select all the relevant synonyms and concepts for describing or overlapping with the term AI. Accordingly, the chosen keywords were artificial intelligence, AI, deep learning, machine learning and black box, algorithm and algorithmic. All the keywords were combined with the terms auditing and audit. Terminology in AI varies considerably, and different researchers tend to use different words for AI or ML systems (Ongsulee et al. 2017) which is why all of these were taken into this search. Search strings and keywords are visualized in table 4:

Table 4 Database search syntax

| Data Source | Search Syntax |
|---|---|
| **Scopus** | TITLE-ABS-KEY ( "artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "black box" OR “algorithm” AND "auditing" ) AND ( LIMIT-TO ( SRCTYPE , "j" ) OR LIMIT-TO ( SRCTYPE , "p" ) OR LIMIT-TO ( SRCTYPE , "d" ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" ) OR LIMIT-TO ( DOCTYPE , "cr" ) OR LIMIT-TO ( DOCTYPE , "sh" ) OR LIMIT-TO ( DOCTYPE , "no" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) ) |
| **Web of Science Core collection** | (TS=("artificial intelligence" AND "auditing") OR TS= ("AI" AND "auditing") OR TS=("machine learning" AND "auditing") OR TS=("deep learning" AND "auditing") OR TS=("black box" AND "auditing" OR TS=(“algorithmic auditing”) OR TS=(“algorithm auditing”)) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Abstract of Published Item OR Data Paper OR Discussion OR Proceedings Paper OR Reprint OR Review) |
| **IEEE Xplore** | ("Full Text & Metadata": artificial intelligence AND auditing OR ai AND auditing OR machine learning AND auditing OR deep learning AND auditing OR black box AND auditing OR “algorithmic auditing” OR “algorithm auditing”))<br>Filters Applied:<br>Conferences Journals Magazines Early Access Articles |
| **ACM Digital Library conference proceedings** | "query": { AllField:("machine learning" AND "auditing" OR "AI" AND "auditing" OR "deep learning" AND "auditing" OR "black box" AND "auditing" OR "artificial intelligence" AND "auditing" OR “algorithmic auditing” OR “algorithm auditing”)) }<br>"filter": { ACM Pub type: Proceedings, Journals, Article Type: Research Article } |
| **ACM Digital Library journal publications** | "query": { AllField:("machine learning" AND "auditing" OR "AI" AND "auditing" OR "deep learning" AND "auditing" OR "black box" AND "auditing" OR "artificial intelligence" AND "auditing" OR “algorithmic auditing” OR “algorithm auditing”) }<br>"filter": { ACM Pub type: Journals } |

The database search resulted in a total of 58 articles. In addition, I conducted backward citation chaining for this sample. Backward citation search is a search to find all the cited references in a single article. It shows what led to the article and is a good way to ensure that no relevant articles are missing (Hu et al. 2011). The reference list from each article was screened based on titles. Full text review was conducted for articles which seemed relevant based on titles. Backward citation chaining resulted in a total of 30 articles which were included in the final sample.

There is no individual reason why backward citation chaining articles did not appear when screening databases. Human error is of course one possible reason. It is possible that during the screening phase I missed articles which appeared again when doing backward citation chaining. Another possible reason is that articles were not in selected databases. This is also the reason why backward citation chaining is made. I mentioned earlier that the selected four databases were enough to cover the field of AI auditing and backward citation chaining further increases the accuracy. Backward citation chaining added to the sample from databases and resulted in a total of 88 articles. This was the final sample that I set out to analyze.

There is an increasing popularity for including grey literature in systematic reviews. This is because relevant frameworks might be missing, if the search is targeted only for academic papers. Including grey literature can broaden the scope of the review by providing a more complete view of the field. (Mahood et al. 2013.) Different definitions of grey literature have been made. One of the most common is the definition provide by Grey Literature Network Service:

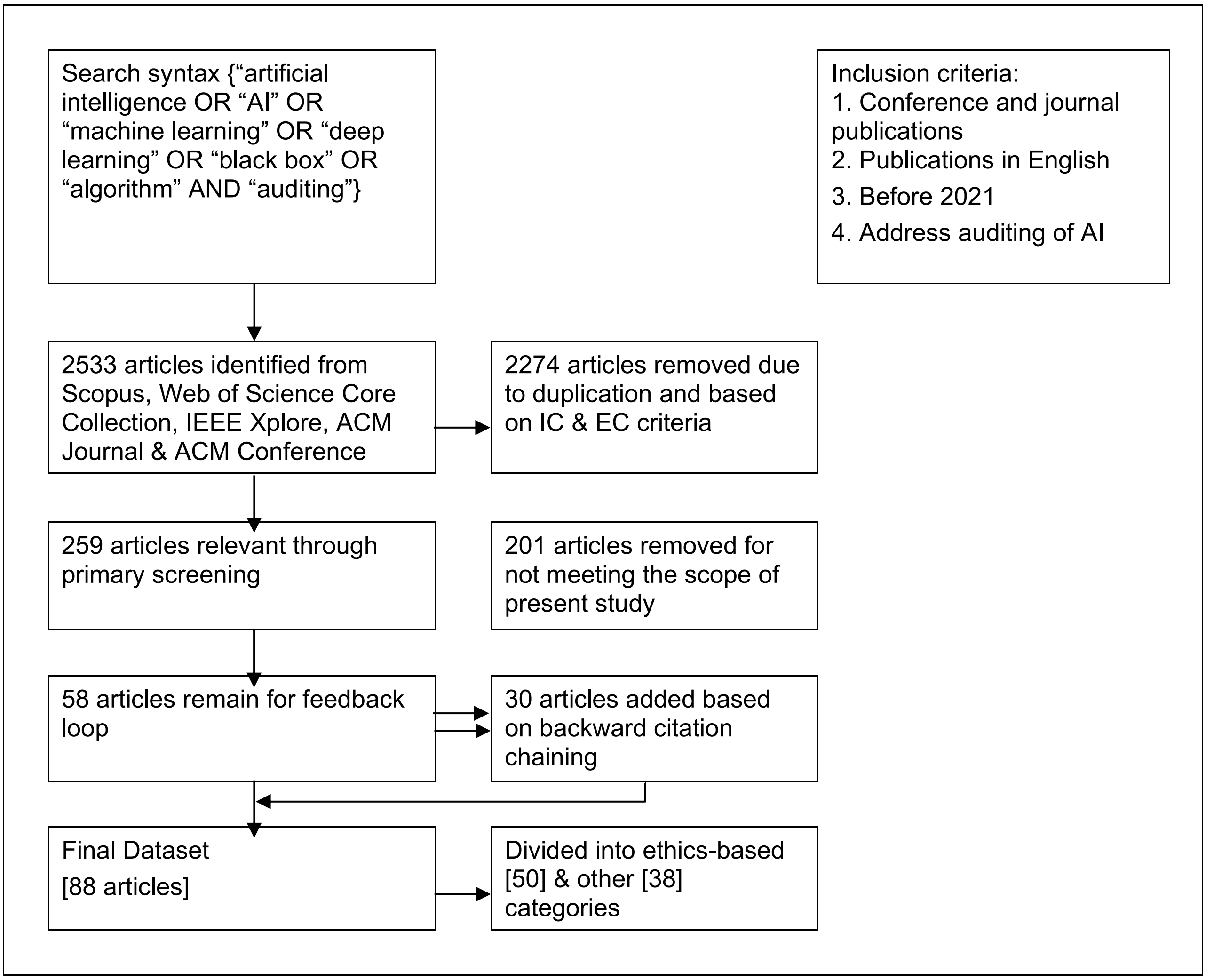
*Grey Literature is a field in library and information science that deals with the production, distribution, and access to multiple document types produced on all levels of government, academics, business, and organization in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body (GreyNet 2013).*

However, due to the relatively large sample size from the database search and the academic nature of this thesis, this study does not include grey literature. It is worth noting that academic literature might come a little bit slower than grey literature. This is because big technology companies have their own interest in developing AI auditing, and they might be doing their own publications. For example, KPMG published their own risk and control framework designated to guide AI professionals (KPMG 2018). Also, AI field is still new and the newer the field is, the larger is the amount of the grey literature. One reason might also be the open source in information systems science. Someone just advertises their projects or finding and reports it, for example, in their own websites while academic literature has its own procedure before publications.

Therefore, backward citation chaining was the only complementary search for the database searches. I screened manually all the reference lists of the articles in order to identify relevant articles. Total of 30 additional articles were found during backward citation chaining. For the best practices, I also exhausted citation chaining within these thirty articles, but no more relevant articles were found. Therefore, final sample consisted in a total of 88 articles which was later divided into ethics-based and other category articles.

I divided the original sample into ethics-based and other category papers. As discussed in section 1.4, papers were defined as ethics-based papers if they matched criteria of AI HLEG (2019). Division between ethics-based and other category papers was based on titles, abstracts and keywords. If a paper matched AI HLEG (2019) criteria of ethics guidelines, it was labeled as an ethics-based paper. This resulted in a total of 50 ethical papers for further analysis where the baseline was based on AI ethical principles presented by Jobin et. al (2019).

Table 5 Flowchart of the search process


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 2021
4. Address auditing of AI
2533 articles identified from Scopus, Web of Science Core Collection, IEEE Xplore, ACM Journal & ACM Conference
2274 articles removed due to duplication and based on IC & EC criteria
259 articles relevant through primary screening
201 articles removed for not meeting the scope of present study
58 articles remain for feedback loop
30 articles added based on backward citation chaining
Final Dataset
[88 articles]
Divided into ethics-based [50] & other [38] categories

## 2.3 Data extraction

The final sample consisted of 50 papers, based on which the data analysis was conducted. First, I collected descriptive data from every paper. That contained publication years, division between journal and conference publications, research methods and relevant details. In addition, I identified key objectives, finding and problems mentioned in papers. This helped to map the overall picture of the field as well as the direction and trends where it is going and what has been done.

Second, sample documents were screened for identifying ethical AI principles identified by Jobin et al. (2019). The goal was to sort out how the ethical AI principles are seen in the ethics-based AI auditing literature, what the most important principles are and how those principles link to the AI auditing. This way key drivers and dimensions of ethics-based AI auditing can be recognized and analyzed. Ethical principles were divided between different categories based on ethical concepts. Each concept included several principles. Separately each concept was analyzed to find out how it drives auditing AI and on the other hand, in what ways it is important for auditing AI.

Last, I identified auditing targets. This means stakeholders of AI auditing, who are proposed to conduct the audits, what aspects are important when considering carrying out AI auditing and how it affects certain groups. Identified AI auditing conductors or targets for AI auditing were researchers, system developers and deployers, regulators, individuals and companies, auditors and users. These actors were then further analyzed to identify their roles, responsibilities, benefits or risks of auditing AI.

# 3 Findings

This section presents the findings from the review. Overall, I identified 50 studies matching the inclusion and exclusion criteria and whose findings addressed research questions. Section 3.1 presents the descriptive details of these articles. Section 3.2 presents the AI auditing ethical principles. I will go through drivers and dimensions in a cross-tabulation format and present main findings on each of them. Section 3.3 describe the main results of the actor-based approach, answering questions about who the stakeholders of AI auditing are, who conducts AI auditing and what problems auditing is trying to solve or bring up.

## 3.1 Descriptive details

Table 6 Sample from the databases

| STUDY | ARTICLE | METHOD | DETAILS |
|---|---|---|---|
| **P1** | Raji et al. 2020 | Design science | SMACTR framework |
| **P2** | LaBrie et al. 2019 | Design science | An Ethical AI Algorithm Audit Framework |
| **P3** | Malgieri & Comande 2017 | Conceptual & desing science | The legibility test |
| **P4** | Barlas et al. 2019 | Empirical | The Social B(eye)as Dataset |
| **P5** | Kearns et al. 2019 | Empirical | Empirical investigation of the SUBGROUP algorithm on four datasets |
| **P6** | Raji & Buolamwini 2019 | Empirical | Modeled Gender Shades |
| **P7** | Kim et al. 2019 | Design science & empirical | Framework of multiaccuracy auditing |
| **P8** | Domingo-Ferrer et al. 2019 | Empirical | Collaborative rule-based model |
| **P9** | Sulaimon et al. 2019 | Design science | Control Loop framework |
| **P10** | Martinez & Fernandez 2019 | Conceptual | A Multi-agent system architecture |
| **P11** | Cabrera et al. 2019 | Design science | FAIRVIS |
| **P12** | Singh et al. 2019 | Conceptual | Decision provenance |
| **P13** | Raji et al. 2020 | Empirical & Design science | CelebSET |
| **P14** | Clavell et al. 2020 | Empirical | The Algorithmic Audit of REM!X |
| **P15** | Dulhanty et al. 2020 | Empirical | MS-Celeb-1M |
| **P16** | Harrison et al. 2020 | Empirical | Online between-subject, survey-based experiment |

| | | | |
|---|---|---|---|
| **P17** | Papakyriakopoulos et al. 2020 | Design science | A new technique for bias detection |
| **P18** | Black et al. 2020 | Design science | FlipTest |
| **P19** | Ilvento et al. 2020 | Design science | A stylized model and fairness requirements that match the intuitive fairness desiderata |
| **P20** | Katell et al. 2020 | Design science | Algorithmic Equity Toolkit |
| **P21** | D’Amour et al. 2020 | Design science | An extensible open-source software framework |
| **P22** | Singh & Hofenbitzer 2019 | Empirical | Labeled cyberbullying dataset |
| **P23** | Barlas et al. 2019 | Empirical | A between-subjects experiment at MTurk |
| **P24** | Sapiezynski et al. 2019 | Design science | The Viable-Λ test |
| **P25** | Chen et al. 2018 | Empirical | Controlled experiment of 855K job candidates |
| **P26** | Mehrotra et al. 2019 | Empirical & design science | A framework for internally auditing online services |
| **P27** | Kulshrestha et al. 2017 | Design science | A framework for to quantify ranking systems biases |
| **P28** | Robertson et al. 2018 | Empirical | Survey for 187 participants |
| **P36** | Brown et al. 2020 | Design science | Auditing framework to guide ethical assessment of an algorithm |
| **P37** | Toapanta et al. 2020 | Empirical & design science | Developed a prototype which performs audits on social networks |
| **P38** | Shneiderman 2020 | Conceptual | 15 recommendations at three levels of governance |
| **P39** | Barlas et al. 2020 | Empirical | Controlled experiment on the interdependence between algorithm recognition and persons’ gender |
| **P40** | Scheuerman et al. 2020 | Empirical | Sample of 92 image databases |
| **P41** | Grasso et al. 2020 | Empirical | Algorithmic accountability framework |
| **P42** | Jiang & Vosoughi 2020 | Empirical | Statistical differences in performances of MTurk annotators |

Table 6 presents the full sample from the databases, highlighting the research method and relevant details. A total of 35 articles were included to the further analysis.

Table 7 Sample from the citation chaining

| STUDY | ARTICLE | METHOD | DETAILS |
|---|---|---|---|
| **P29** | Buolamwini & Gebru 2018 | Empirical | An approach to evaluate bias in automated facial analysis algorithms and datasets |
| **P30** | Kroll et al. 2016 | Conceptual | A new technological toolkit for automated decisions and standards of legal fairness |
| **P31** | Sandvig et al. 2014 | Conceptual & empirical | Outlined five idealized audit designs for empirical research projects investigating algorithms |
| **P32** | Tan et al. 2018 | Empirical & design science | Transparent model distillation approach |
| **P33** | Singh et al. 2016 | Conceptual | Discussion paper about responsibility in ML |
| **P34** | Reed et al. 2016 | Conceptual | Investigation of legal liability in ML |
| **P35** | Goodman 2016 | Conceptual | Investigation about how EU GDPR address discrimination |
| **P43** | Floridi et al. 2018 | Conceptual | AI4People—An Ethical Framework for a Good AI Society |
| **P44** | Mittelstadt 2019 | Conceptual | Critical assess of the strategies and recommendations proposed by current AI Ethics initiatives |
| **P45** | Mittelstadt et al. 2016 | Conceptual | Conceptual framework aiming to inform future ethical inquiry, development, and governance of algorithms |
| **P46** | Obermeyer et al. 2019 | Empirical | Series of experiments for the sample of 6079 patiens |
| **P47** | Kyriakou et al. 2018 | Empirical | A set of descriptive tags for all images in the Chicago Face Database, using the six tagging APIs. |
| **P48** | Bellamy et al. 2018 | Design science | AI Fairness 360 Toolkit |
| **P49** | Hanna et al. 2019 | Conceptual | Ground conceptualizations of race for fairness research, |
| **P50** | Mehrabi et al. 2019 | Empirical | A Survey on Bias and Fairness in Machine Learning |

Table 7 presents the articles from the backward citation chaining, research methods and details from the articles. A total of 15 articles were included to the further analysis from the backward citation chaining.

From a sample of 50 articles, 35 were conference papers and 15 were journal articles. All of the studies were published between 2016 and 2020. As auditing of artificial intelligence is a relatively new topic, it can clearly be seen that most of the studies are made in 2019 or 2020. Figure 2 shows the number of journal publications and conference proceedings of each year and figure 3 shows in which journal or conference papers were published. The complete list of the articles is given in tables 6 and 7.

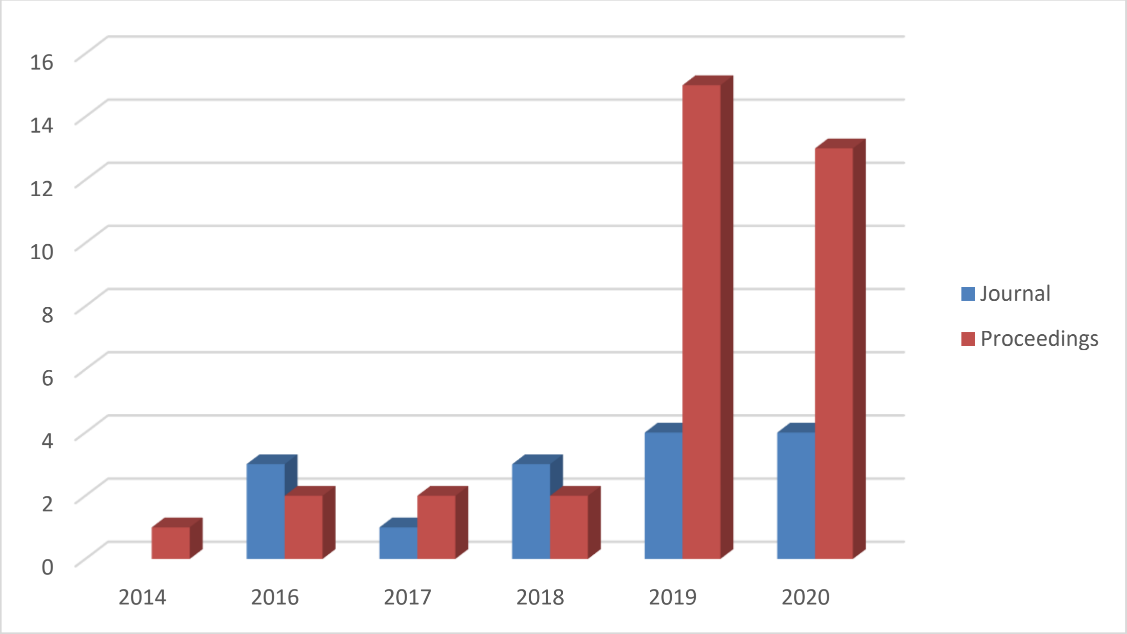


Figure 2 Publications of each year

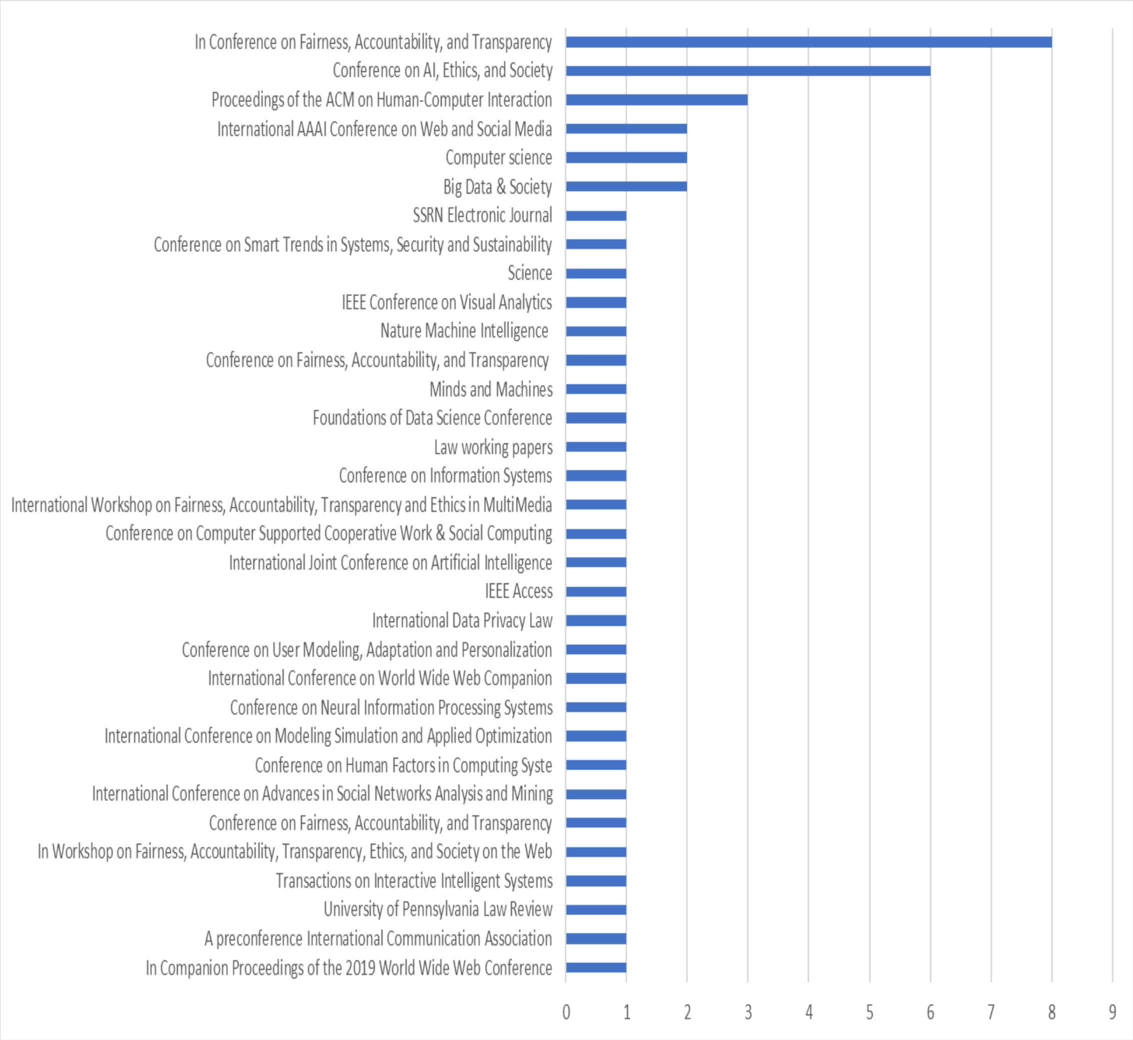


Figure 3 Journals and conferences

In particular, the large numbers of conference papers stand out clearly which shows in Figure 2. The reason is most likely that in information systems science and AI literature most people prefer conference publications. The topic is uncertain and difficult to understand so researchers benefit having a discussion with others. Many people attend researchers' seminars, and they have opportunities to ask questions which will benefit in both ways. In the IT field, conference papers might also provide higher visibility and greater impact. High status conferences will probably get more audience. In addition, it is faster to publish conference papers. AI auditing field is moving so fast that usually the papers need to be out as fast as possible.

As Figure 3 presents, three conferences were clearly ahead of others. Conference on Fairness, Accountability, and Transparency had eight hits, Conference on AI, Ethics, and society had six hits and Proceedings of the ACM on Human-Computer Interaction had three hits. The rest of the conferences or journals had just one or two mentions. Based on the titles of these journals and conferences, these top three conferences also seem to be most relevant for the given topic. Themes correspond with the demand when authors are writing about ethics-based AI or AI auditing. Especially Conference of Fairness, Accountability, and Transparency is a key forum for the discussion of ethics-based AI auditing.

The sample was relatively evenly distributed between the research methods. 11 studies were recognized as conceptual studies, 13 as design science and 19 as empirical studies. The other 7 studies used two of those methods. The original sample was divided into papers focusing on ethics-based principles and papers focusing on something else, therefore, the problems that the papers approached were relatively similar. The approach to the problems depended on the research method.

Studies using design science seek to develop a framework or artifacts for certain ethical problems. For example, Raji et al. (2020) created the SMACTR-framework for identifying the harmful repercussions of systems prior to deployment and after deployment, Cabera et al. (2019) created the FAIRVIS-framework for discovering which biases machine learning models has introduced and Black et al. (2020) created the FlipTest-framework, a black-box technique, for uncovering discrimination in classifiers. Also, design science studies which did not create their own framework tried to seek solutions for certain ethical problems. Bellamy et al. (2018), Ilvento et al. (2020) and Sulaimon et al. (2019) created designs for improving fairness in the decision-making processes of autonomous software systems. In addition, Sutton & Samavi (2019) combined the PAPA-framework and the layered model of AI governance to demonstrate that our algorithms might be biased or incomplete. Algorithm biases were also the concern of Kulshrestha et al. (2017), Kim et al. (2019) and Papakyriakopoulos et al. (2020).

Experiments were the most common method for conducting empirical studies. Eight studies used experiment as their research method. The common factor in the experiment studies was that they tried to prove or demonstrate problems like struggle with gender recognition (Barlas et al. 2020), discrimination based on race or gender (Buolamwini & Gebru 2018) or lack of auditing models on social network (Toapanta et al. 2020). Case study was the second most used method. As in experiments, through case studies researchers aim to bring attention to ethical problems such as lack of transparency, accountability, and fairness (Grasso et al. 2020) and biases and unfairness in online platforms (Mehrotra et al. 2019; Robertson et al. 2018).

Conceptual papers were more solution-oriented or descriptive of the technology systems. Some of the studies were very practical, like Martinez & Fernandez (2019) whose paper discussed about artificial intelligence in recruiting and challenges behind the analysis of job video interviews. Singh et al. (2016), Mittelstadt et al. (2016) and Singh et al. (2019) are examples of more technical papers. They addressed the impact of machine learning systems and how those can be controlled, how data should be interpreted, what actions should be taken and what accountability and fairness challenges artificial intelligence systems face.

## 3.2 Ethical principles in AI auditing

As was already established above, the field of ethics-based AI auditing is relatively new and unsettled. Therefore, it is important to take a closer look at the drivers and dimensions which are leading the field right now. ‘Driver’ describes what and why something is happening in the field and ‘dimension’ on the other hand tells how something is happening. This section focuses on how much and in which way ethical principles identified in existing AI guidelines by Jobin et al. (2019) appear in existing ethics-based AI auditing literature.

Table 8 Research method & ethical principles cross-table

| | ***Design science*** | ***Conceptual*** | ***Empirical*** |
|---|---|---|---|
| ***Transparency*** | P1, P13, P18, P20, P32, P36, P48 | P3, P12, P30, P31, P33, P34, P35, P38, P43, P45 | P6, P8, P14, P23, P25, P29, P39, P41, P42, P47, P50 |
| ***Justice & fairness*** | P1, P7, P9, P10, P11, P13, P17, P18, P19, P20, P21, P24, P27, P32, P36, P48 | P2, P3 P12, P30, P31, P34. P35, P38, P43, P45, P49 | P4, P5, P6, P8, P14, P15, P16, P22, P23, P25, P26, P28, P29, P39, P41, P42, P46, P47, P50 |
| ***Non-maleficence*** | P1, P7, P13, P20, P21, P36 | P2, P3, P12, P30, P31, P33, P34, P35, P38, P43, P44 | P6, P14, P37, P42, P50 |
| ***Responsibility*** | P1, P13, P20 | P12, P30, P31, P33, P34, P35, P38, P43, P44 | P6, P14, P16, P23, P29, P41, P42 |

| | | | |
|---|---|---|---|
| ***Privacy*** | P1, P13 | P2, P3, P12, P30, P33 | P6, P8 |
| ***Beneficence*** | P10, P48 | P3, P12, P30, P34, P38, P43 | P8 |
| ***Freedom and autonomy*** | P13 | P3, P30, P34, P43 | P15 |
| ***Trust*** | P36 | P12, P30, P38, P43 | P16, P28 |
| ***Sustainability*** | | | |
| ***Dignity*** | | | |
| ***Solidarity*** | | | |

Table 8 shows how different ethical principles were shown in papers. The term had to be linked with AI auditing for matching the criteria. The top row also shows the research method. Paper codes are presented in table 6 and table 7. More detailed codes for the principles can be seen in table 9.

Table 9 Ethical principles identified in existing AI auditing guidelines (Jobin et al. 2019)

| ***Ethical principle*** | ***Number of documents*** | ***Included codes*** |
|---|---|---|
| *Justice & fairness* | 46 | Justice, fairness, consistency, inclusion, equality, equity, bias, discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access, distribution |
| *Transparency* | 28 | Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| *Responsibility* | 19 | Responsibility, accountability, liability, acting with integrity |
| *Non-maleficence* | 22 | Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity, non-subversion |
| *Privacy* | 9 | Privacy, personal or private information |
| *Beneficence* | 9 | Benefits, beneficence, well-being, peace, social good, common good |
| *Trust* | 7 | Trust |
| *Freedom & autonomy* | 6 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |

Table 9 presents the ethical principles by Jobin et al. (2019), number of documents for each principle in the sample and codes included in the principles. Following sections analyzes further these principles and how they link with ethics-based AI auditing. As we can see, justice and fairness following with transparency, responsibility and non-maleficence were the most common principles. Literature also identified privacy, beneficence, trust and freedom and autonomy, but to a smaller extent. It is worth mentioning that Jobin et al. (2019) principles also identified sustainability, dignity, and environmental themes in AI ethics, but the principles were not found in the AI auditing sample.

### 3.2.1 Fairness & Justice

Featured in 46 of our 50 sources, fairness and justice is by far the most prevalent principle in the AI auditing literature. As presented earlier, justice and fairness codes included terms justice, fairness, consistency, inclusion, equality, bias, discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution (Jobin et al. 2019). Of these, fairness and bias were the most popular principles. Fairness occurred in headlines 13 times while bias occurred 11 times. The majority of other papers included also fairness and bias as the most dominant terms in this category. For the sake of clarity, in the following texts the word 'fairness' does not include all the terms above but deals only with the term fairness.

Fairness is a diverse concept. Cabrera et al. (2019), Clavell et al. (2020), Harrison et al. (2020), Papakyriakopoulos et al. (2020), Kroll et al. (2016), Grasso et al. (2020) and Bellamy et al. (2018) all highlight multiple definitions of fairness and challenges due that. No unique fairness definition exists, and researchers need to point out in what perspective they consider fairness. 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 (Grasso et al. 2020; Bellamy et al. 2018) while Barlas et al. (2019) note that fairness "*is best understood as a placeholder term for a variety of normative egalitarian considerations*." Overall, 41 sources recognized the word in their paper.

Fairness was mostly presented as a human identity issue or when a specific group or an individual receive unfavorable treatment (Raji & Buolamwini 2019; Sulaimon et al. 2019; Dulhanty & Wong 2020; Papakyriakopoulos et al. 2020; Singh & Hofenbitzer 2019 Sapiezynski et al. 2019). Race and gender have become two of the major concerns regarding bias in machine learning fairness literature, and fairness is to treat subjects similarly regardless of their defined protected attributes (Sulaimon et al. 2019; Scheuerman et al. 2020). In the decision-making process, fairness is the absence of any preconception, discrimination or favor toward an individual or a group based on their inherent or acquired characteristics (Mehrabi et al. 2019).

An important concept was also making a difference between group fairness and individual fairness (Kearns et al. 2019; Raji & Buolamwini 2019; Kim et al. 2019; Cabrera et al. 2019; Raji et al. 2020; Clavell et al. 2020; Papakyriakopoulos et al. 2020; Black et al. 2020; Sapiezynski et al. 2019; Chen et al. 2018; Barlas et al. 2020; Bellamy et al. 2018; Hanna et al. 2019). 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*" while individual fairness is "*the goal of similar individuals receiving similar treatments or outcomes*" (Bellamy et al. 2018). Audits of algorithms systems typically highlight the notion of group fairness, which holds that advantaged and protected groups should be treated same way than others. In comparison, individual fairness deal with the notion of consistency, holding that systems should treat similar individuals in the same way. (Barlas et al. 2020.)

Kearns et al. (2019) found out in their study that even algorithms which are explicitly equalizing false positive rates across the groups defined on the marginal protected attributes often violates subgroup fairness constraints. This is where AI auditing is needed, and for example, Harrison et al. (2020) state that to audit the model or to achieve group fairness race may be needed to be taken into consideration. Kim et al. (2019) and Barlas et al. (2019) used auditing for determining the same effect of subgroup fairness and whether the predictor satisfies a strong notion of subgroup fairness, multiaccuracy. Multiaccuracy requires that predictions are fair and unbiased, and that researchers are developing auditing processes to make algorithms more transparent and fairer. They want to open the black box. Same goes with Sapiezynski et al. (2019) as they investigate fairness problems in black-box systems from the algorithmic auditing side.

Researchers and auditors have made an increasing number of strategies for testing and detecting discriminatory behavior and unfairness. The problem is that certain models can pass such audits while behaving in an unfair way. This is why Black et al. (2020) recommend a setting where audits are conducted towards a ML system by the stakeholders who have been involved in the model's construction or practitioners outside of the development process and Ilvento et al. (2020) state that to fulfill platform fairness requirements auditing needs to be outsourced to a neutral third-party or governing body. Algorithmic auditing community wants to measure that models operate in a fair and unbiased way, but challenge is to select appropriate metrics for assessing results. For example, auditing real-world search engines metrics could be an average representation that fail to take order effect into account, group representation in top ranks, logarithmic discounting and linear normalization by rank to model the decay of attention. These methods could lead to incorrect conclusions about fairness. (Sapiezynski et al. 2019.)

Kyriakou et al. (2018) and Hanna et al. (2019) studied fairness in AI auditing from image tagging and race perspective. This racial bias concept, as with gender bias, is one of the most common topics and it is often very close matter to the fairness. Only three studies out of 50 did not recognize the word bias in their work. Clavell et al. (2020) define bias as unfavorable treatment of an already disadvantaged group. They also state that the criteria by which something constitutes as bias should be framed from a social and ethical standpoint as some features may be legal or hold legitimate for unfavorable treatment, but still seen as discriminatory in some contexts or reasons.

Numerous definitions of fairness and bias has created a challenge for discovering biases in ML models. Naturally encoded societal biases in the ML models are often referred as algorithmic biases. These algorithmic biases should be addressed before deploying ML systems which is why it is vital to audit ML models. However, recognizing biases can be hard due to the inherent intersectionality of bias. Intersectional bias means that populations are defined by multiple features. Auditing bias can be straightforward when having only a few features and a single definition of fairness, but with increasing number of features the number of populations grows and quickly becomes unmanageable. The visal analytics system, FAIRVIS, was developed for discovering intersectional bias and help data scientists better audit their models. (Cabrera et al. 2019.)

Bias detection mechanisms are made for ensuring fairness in the decision-making processes (Sulaimon et al. 2019). In addition, these mechanisms are needed for detecting bias for gendered languages to compare bias in embeddings trained on social media data (Papakyriakopoulos et al. 2020). Bias shares public opinions, for example, in political events which is why it is needed to distinguish and quantify between the bias that arises from the data that serves as the input to the ranking system and the bias that arises from the ranking system itself. Both produce varying amounts of bias, and the consequences of these biases needs to be addressed. (Kulshrestha et al. 2017.) Also, Google search results and personalization within Google search indicate the need for an audit. The participant bias can significantly influence public opinions and affect, for example, voting intentions. Controlled algorithm audit is thus needed to assess the partisan audience bias. (Robertson et al. 2018.)

### 3.2.2 Transparency

The second most prevalent dimension of AI auditing is transparency. A total of 28 papers featured transparency in their studies. Other than transparency, codes also included explainability, explicability, understandability, interpretability, communication, disclosure and showing (Jobin et al. 2019). Transparency appeared in various different contexts, primarily linked with general data ethics principles as a way to be operationalized via algorithmic auditing to lead better technologies (Raji et al. 2020; Clavell et al. 2020; Black et al. 2020; Sandvig et al. 2014; Reed et al. 2016; Barlas et al. 2020; Grasso et al. 2020), as a way to minimize harm and improve AI (Domingo-Ferrer et al. 2019; Raji & Buolamwini 2019; Barlas et al. 2019; Katell et al. 2020; Buolamwini & Gebru 2018; Tan et al. 2018; Brown et al; 2020, Jiang & Vosoughi 2020; Kyriakou et al. 2018) or as a way to improve responsibility and explainability issues (Malgieri & Comande 2017; Singh et al. 2019; Singh et al. 2016; Goodman 2016; Shneiderman 2020; Floridi et al. 2018; Mittelstadt et al. 2016). According to Mittelstadt et al. (2016), transparency is generally defined as "*the availability of information, the conditions of accessibility and how the information may pragmatically or epistemically support the user's decision-making process*". Buolamwini & Gebru (2018) added human-centered vision where transparency is defined as a demographic and phenotypic composition of training and benchmark datasets. Transparency is also often presented as a close principle with fairness and accountability (Raji et al. 2020; Singh et al. 2019; Katell et al. 2020; Barlas et al. 2019; Grasso et al. 2020).

Barlas et al. (2020) & Kyriakou et al. (2018) note that for an ML system to be transparent, this requires that algorithmic tools must be open, and users and developers must have the skills to understand them. Open tools mean that trade secrets cannot be protected, and the system is interpretable from a technical point of view. Systems in which full transparency is impossible, auditing the algorithms should happen from the outside. In this regard, Raji et al. (2020) state that people, organizations or other audit targets rather dismiss than act on results if any dishonesty or non-transparency happens in audit methodology, therefore, auditors need to live up with high ethical standards. Internal audits generate transparent information and artifacts and complement external accountability so that third parties can use external auditing.

While accountability focuses on methods for holding a system to an ethical standard determined by domain experts, transparency refers to understanding the inner mechanisms of why certain outputs comes out from algorithms. (Grasso et al. 2020.) Reed et al. (2016) note that transparency is a wider concept than accountability, being a property of a system providing visibility of its governing norms and behavior. Inner mechanisms are exposing the critical knobs of the decision-making process which later helps developers apply a code of ethics in ML systems (Grasso et al. 2020). Via a transparency report, it is possible to gain information about why the model behaves in a certain way. For example, transparency reports can identify features that are responsible for the model's bias. (Black et al. 2020.) Transparency reports require reasons why and how ML technology makes decisions, and it is the most important accountability attribute for liability questions (Reed et al. 2016).

Auditing processes are being developed for making algorithms more transparent for users and promoting fairness. This is called "opening the black box". (Barlas et al. 2019.) When it comes to direct corporations towards transparency, accountability or fairness, external pressure remains necessary as they hesitate to disclose details about their systems (Raji & Buolamwini 2019) which is why Kyriakou et al. (2018) note that different associations, IEEE and ACM for example, are encouraging developers to take measures to promote transparency in the algorithmic systems they build. Public scandals have increased the ethical impact in AI systems, and lack of transparency or misuse of data have often been key issues in those scandals. That is why audits focus primarily on issues of transparency and autonomy. Transparency of architecture measures how well stakeholders know the structure of the algorithms, transparency of use measures how algorithms are being used and transparency of data & use measures how well the collection and subsequent use of data for the algorithm are known to stakeholders. AI auditing secures that potential abuses and misuses are reduced, legal rights are not infringement, security and access of use are not violated and data is secured. (Brown et al. 2020.)

Transparency enables identification, audit and oversight which in turn help holding systems responsible. Technical measures can support in making systems more transparent and transparency is often a regulatory requirement for identifying responsibility. (Singh et al. 2019.) Shneiderman (2020) agrees with Singh et al. (2019) arguing that transparency enhances correctness, identifies improvements, accounts for changing realities, supports users in taking control and increases user acceptance. Poorly designed algorithms are hard to control and monitor which is why transparency is needed. Poor transparency can also harm other ethical ideas, privacy of data subject and autonomy in particular. Key components of transparency, accessibility and comprehensibility, are for making sure that needed information about algorithms is accessible and comprehensible. Auditing is necessary to verify correct functioning. It can, for example, make a path to explainability by making a record of complex algorithmic decision-making to unpack problematic or inaccurate decisions. (Mittelstadt et al. 2016.)

Malgieri & Comande (2017) present algorithm legibility to combine comprehensibility into transparency. It is a fundamental tool to empower data subject in the algorithmic era. The purpose of legibility is to let individuals autonomously understand the functionality, the impact, the consequences and the rationales of decision-making processes. Auditing of decision-making algorithms can identify ML bias issues which can expose liability and sanctions. These liability issues are also closely connected with Singh et al. (2016) research of control and transparency.

### 3.2.3 Non-maleficence

References to non-maleficence encompass mostly calls that AI should not cause any harm. Codes also include non-maleficence, security, safety, protection, precaution, prevention, integrity and non-subversion (Jobin et al. 2019) which featured a total of 21 papers. Clear definition for non-maleficence concepts did not appear in the literature but codes mostly addressed or filled other dimensions like fairness and bias issues (Raji et al. 2020; LaBrie et al. 2019; Raji & Buolamwini 2019; Kim et al. 2019; Raji et al. 2020; Clavell et al. 2020; Sandvig et al. 2014; Goodman 2016) or accountability and liability issues (Malgieri et al. 2017; Singh et al. 2019; Singh et al. 2016), or they pointed out potential harms (D'Amour et al. 2020; Reed et al. 2016; Brown et al. 2020; Shneiderman 2020; Jiang & Vosoughi 2020; Mittelstadt 2019; Mehrabi et al. 2019). LaBrie et al. (2019) and Floridi et al. (2018) equated non-maleficence with beneficence as beneficence is based on doing only good and non-maleficence is based on doing no harm. Floridi et al. (2018) also defined justice as preventing the creation of new harms, such as the undermining of existing social structures and ensuring that AI creates benefits and eliminates unfair discrimination.

One of the most important reasons for auditing is to identify harmful repercussions and to prevent those. However, it is important to separate system reliability harms from the societal harms. Raji et al. (2020) argue that an AI system might be technically reliable but does not meet ethical expectations. Potential sources of harm and social impacts are then screened through auditing. One method Raji et al. (2020) suggest to identify harms which are caused by AI systems is social impact assessment. It is for analyzing and mitigating the unintended social consequences with two primary steps: assessments of risks and identification of the relevant impacts and harms that are caused by AI systems.

According to LaBrie et al. (2019), ethical AI auditing is for providing external information about doing no harm, meaning detecting and calling out potential biases, harms or flaws. Therefore, the goal for auditing is to minimize harmful biases. Kim et al. (2019) share this “do no harm” view while searching systematic biases which harm specific subgroups. As AI systems become more widespread, the external pressure for addressing harmful biases increases (Raji & Buolamwini 2019). Marginalized populations need to be protected and ethical guidelines, policies and corporate practices are needed for ensuring that evolving AI technology does not cause harm. This is also addressed by Sandvig et al. (2014) and Goodman (2016) who highlight the need of audit studies for diagnosing harmful discrimination. Especially in companies with big data repositories, like Facebook, YouTube or Google, it is important to investigate the operation of their algorithms consequences whether they are conducting harmful discrimination.

Accountability links with harm due liability. Accountability involves determining liability for an action and if harm arises it determines what restitution is owed by who and to whom for that harm. Decision provenance is one method for identifying causes of harms when it is caused by failure by recognizing responsible and liable actors and making them accountable. (Singh et al. 2019.) Singh et al. (2016) also note that responsibility generally leads to liability, and it helps to identify harms and persons or organizations causing it. Therefore, as autonomous systems have the potential to cause harm, responsibility is needed for holding persons managing systems accountable which will be addressed in section 3.2.4.

Potential of different harms growths with greater capabilities of technology. One goal with AI auditing is to reduce potential of unexpected harmful outcomes with human-centered AI systems and make sure that systems do what users expect. For example, robotic devices could cause multiple different safety problems. Improved safety is linked with decreased harm. Certain contracts even contain "hold harmless" clauses which releases developers from their liability. Human-centered AI systems (HCAI) movement linking with AI auditing raise these issues with calls of accountability and transparency. HCAI focuses on creating reliable safe and trustworthy systems by amplifying, augmenting, and enhancing human performance while traditional AI science focuses on emulating human behavior or replacing human performance (Shneiderman 2020). Auditing helps to gain insight into how a technical system performs with multiple indicators. Different toolkits, for example AIF360 and the Perspective API, are developed to survey and to reduce harm caused by AI systems. AI auditing ensures that those toolkits work in the intended way. (Jiang & Vosoughi 2020; Mehrabi et al. 2019.)

### 3.2.4 Responsibility

Responsible AI is one of the most discussed themes related to AI (Dignum 2019). However, it is hardly defined in the literature. Mostly, themes around responsibility and accountability include recommendations and takes on what should be better. 19 papers in total recognized this principle which included terms responsibility, accountability, liability, and integrity (Jobin et al. 2019). It appeared eight times in headline level which highlights the importance of this principle. Responsibility links closely with almost every principle as most of the ethical guidelines somehow lean on responsibility. Grasso et al. (2020) note that best practices for algorithmic accountability are not reached yet and standardization and regulation for data and model usage are still under development.

One of the most cited and noticed AI accountability study is Raji et al. (2020) framework for internal algorithmic auditing aiming to close the AI accountability gap. They define accountability as systems state of being responsible and how systems' answer for its behavior and potential impacts. They note that algorithms are not moral or legal agents, therefore algorithms cannot be held accountable. However, governance and auditing structures can be. These structures include designing and deploying algorithms and they should direct and control the whole organization to achieve its core purpose.

Singh et al. (2019) add to that that systems should be able to apportion responsibility and determine who owns what particular occurrence. This way people and organizations as natural and legal persons are accountable for their actions, and they are also accountable for actions of machines and systems under their control. So, accountability includes determining the responsible person or organization, determining the effects of the particular decision or action and from and to whom an explanation is owed for that happening. Grasso et al. (2020) define it in a little more technical way stating that algorithmic accountability is a method for holding a system to an ethical standard determined by domain experts.

Fairness, transparency and responsibility are all very close principles with accountability. Most of the time they are separated (Clavell et al. 2020, Katell et al. 2020, Buolamwini & Gebru 2018, Shneiderman 2020, Grasso et al. 2020) but Reed et al. (2016), for example, says that transparency is an aspect of the wider concept of accountability and identifies transparency, responsiveness, responsibility, remediability and verifiability as five key concepts of accountability. Transparency is said to be the most important accountability attribute due to its liability questions. Responsiveness means that systems, organizations or individuals need input from external stakeholders to take into account and respond for them. This also links to the liability issues. They define responsibility as "*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, remediability to take corrective action and/or provide a remedy for any party harmed in case of failure and verifiability the extent to which it is possible to assess compliance with accountability norms*."

Katell et al. (2020) developed the Algorithmic Equity Toolkit to investigate fairness, accountability and transparency in algorithmic systems. It includes a set of heuristic tools and an interactive demo that helps users in recognizing algorithmic systems, understanding potential algorithmic harms, and holding policymakers accountable to public input. The study found out that non-technical measures are often more powerful than technical steps like design or development. The role of community organizing, public engagement, and policy oversight in addressing system failures might be a better way to achieve greater accountability. Grasso et al. (2020) approach this fairness, transparency and accountability challenge in an entirely different way. They show how high-level technical solutions make more accountable and transparent decision-making systems by complementing algorithmic accountability frameworks with domain level codes of ethics to investigate this. Domain experts can also expose issues in decision-making systems with the help of accountability mechanisms and in that way apply code of ethics for the automated systems (Grasso et al. 2020).

AI systems must have algorithmic accountability so people can undo unintended harms (Shneiderman 2020), and properly designed algorithmic audits are vital for better accountability (Goodman 2016). However, literature still lacks means to identify in which extent data minimization can be detrimental to the accountability efforts like conducting rigorous algorithmic audits. This data minimization is defined by GDPR but applying it requires to look also at other legal principles or the result might be major limitations in reaching the actual accuracy of the system or other ethical concerns. (Clavell et al. 2020.)

According to Mittelstandt (2019), lack of legal accountability mechanism is one of the four weaknesses of a principled approach to AI Ethics. They ask if it is enough to just define good intentions without actually having a complementary punitive mechanisms and governance bodies. Better accountability mechanisms could be adopted but attitudes and the nature of the field is standing against it. For example, AI developers work mostly in private companies and AI development is a unified profession. Therefore, public interest is not high on the priority list. AI also operates in multiple sectors and changes whole industries so there is great number of different benefits and harms which should be considered with new laws, mechanisms and regulations.

This legal point of view was also considered by Singh et al. (2019) and Kroll et al. (2016). Singh et al. (2019) tied accountability with responsibility, liability and transparency linking it also with GDPR regulation. Accountability can determine liability and point out where harms arise and who is responsible for that. Legal requirements extend beyond transparency, particularly around data protection and privacy and GDPR also emphasizes aspects which produce legal or similarly significant effects, with the so-called ‘right to an explanation’. Technological mechanisms can provide the tooling and information to assist and facilitate accountability, but they do not address those legal concerns themselves. It is worth noting that legal AI auditing partly crosses with ethical AI auditing principles even though it is its own category and not the focus of this thesis.

### 3.2.5 Privacy

Privacy discussions include challenges and values of private and personal information (Jobin et al. 2019). Privacy must be protected as everyone has the right to uphold information about themselves and to be protected. A total of nine papers recognized privacy themes as a key principle. Raji et al. (2020), for example, mention privacy as one of the five most identified AI principles in their End-to-End Framework for Internal Algorithmic auditing. From ethical point of view, they saw privacy involving risks in sensitive juvenile data and biometric face data. Privacy was also a key factor in LaBrie et al. (2019) proposed ethical AI algorithm framework which combined PAPA Framework (Privacy, Accuracy, Property and Accessibility) and layered model of AI governance.

Large datasets can present privacy risks for the individuals represented in the dataset. Auditing datasets are meant to look at who will be impacted by the audited technology, so privacy aspect sets a kind of contradictory challenge. Sensitive and biometric information are stored somewhere and there are risks that those datasets can be reached or be accessible beyond the intended auditing purpose. These consent violations are later discussed in the freedom and autonomy section. There are also risks that these privacy violations exploit marginalized groups. It is shown that privacy risk is increased for members of underrepresented groups. (Raji et al. 2020.)

ML is based on making complex models from the data. That is why the privacy aspect needs to be taken seriously. Different privacy techniques are made for managing privacy in data analytics. Apple, for instance, announced their differential privacy technique that regulate statistical queries to balance the utility of the results with the probability of identifying individual records. Privacy work also focuses on cloud computing aiming smaller clouds which could also improve privacy within the system supply-chain. The idea of smaller clouds is to prepare personal clouds and data stores that gives users more control over processing operations and released data. (Singh et al. 2016.)

Similar data protection issues are discussed by Singh et al. (2019) but from a more juridical point of view. Privacy and data protection are the key components of various legal frameworks and part of legal requirements for accountability. GDPR, for example, sets many standards for the privacy and data controllers obligating data controllers to demonstrate compliance with its various requirements. The EU also has the ePrivacy Regulation which focuses on non-personal data, electronic communications data and information relating to end-user equipment. Hopes are that ePrivacy regulation establish liability where harms are caused. (Singh et al. 2019.) However, Malgieri & Comande (2017) note that privacy harms are accused of being blurred and vague which obscure privacy boundaries and hamper the attempt to contextualize discussion within the general legal theory of privacy.

Kroll et al. (2016) link privacy with fairness and discrimination as fairness can be seen as information hiding. Fairness protects the privacy of certain attributes when a fair decision does not allow us to infer the attributes of a decision subject. People often care more about that their information is not used to make decisions or classifying them than they care about that the information is known or shared. Conception of this is called “right to be let alone”.

Data analysis and classification problems and data aggregation and querying problems are also much discussed in the privacy literature. Just as fairness, private information may be a risk in automated decisions when sensitive information is handled. Decisions about individuals may be based on private data or the decision itself might violate privacy. In theory, personal information could be deleted in the data sets or personal information could remove protected attributes from the input data but both approaches fail to provide the protection. (Kroll et al. 2016.) This privacy preservation is approached by Domingo-Ferrer et al. (2019). They presented a methodology aiming to let individual subject on whom automated decisions have been made to elicit in a collaborative and privacy-preserving manner a rule-based approximation of the model underlying the decision algorithm. It tries to contribute to the challenge where a model demands from users that they share their input features and the labels returned by their queries and truthfully shared data causes privacy problems and falsely accuracy problems. It is so called privacy-accuracy trade-off where this methodology tries to bring a solution.

### 3.2.6 Beneficence

Beneficence which includes terms benefits, beneficence, well-being, peace, social good and common good, features in nine papers (Jobin et al. 2019). Beneficence as a concept promoting good is often mentioned but rarely defined. It is mostly linked with other ethical principles bringing perspective how those could bring benefits in the ML context. For example, Singh et al. (2019) identify both technical and legal benefits for system designers, operators, auditors, regulators and end-users which come from improved accountability. However, they mention that delineated benefits can be interrelated since legal investigation, for instance, may involve technical audit and investigation. Beneficence can be seen as an equivalent to the non-maleficence as “do only good” and “do no harm” represent similar values.

Beneficence as a concept brings a good counterbalance in a difficult topic. Discussion is often very risk-weighted in ML related cases. For instance, self-driving vehicles bring many social and technical benefits, but most people see those as a threat even though studies show that they are safer than vehicles driven by humans. Users of ML systems are likely to seek out technology which has good accountability mechanism, and high social benefits. (Reed et al. 2016.) Beneficence might be hard to concretize but it is typically featured at the top of different lists of principles. That is because “*the development of AI should ultimately promote the well-being of all sentient creatures*” and “*we need to prioritize human well-being as an outcome in all system designs*” like it is stated by Montreal and IEEE. This principle is often characterized as common good. (Floridi et al. 2018.)

Many fields, like healthcare, education or security, have high hopes of benefits of AI. Equally, there are predictions of biased decision-making, killer robots, unfair treatment, violations of personal information and so on. It is important to the whole AI concept to highlight also beneficence which comes from AI. While auditing AI mostly focuses on those challenges or problems, it also allows space for the benefits. Shneiderman (2020) notes that new human-centered AI technologies are meant to clarify who takes action and who is responsible for it. They are designed to be reliable, safe and trustworthy. This brings benefits to the individuals, organizations and society. Achieving these benefits and reliable, safe, and trustworthy AI systems requires that concerns about governmental regulation and non-governmental approaches are taken into account. Seven principles that the European effort listed regarding to this are: technical robustness and safety, human agency and oversight, diversity, non-discrimination and fairness, privacy and data governance transparency, environmental and societal well-being and accountability. (Shneiderman 2020.)

Several research organizations have explored beneficence and benefits of auditing AI. For example, auditing could turn beneficial for data controllers with reduced liability risks and improved decision-making. Even GDPR requires positive actions. Auditing algorithms can be beneficial to both data controllers and data subjects. Machine biases which were discussed in earlier section, for example, is one topic where mutual benefits can be achieved. (Malgieri & Comande 2017.) This is why so many research organizations try to explore long-term impact of AI and find ways to enable the benefits of it for the humanity (Shneiderman 2020).

### 3.2.7 Trust

Trust featured in a total of seven papers. Different scandals over biased outcomes, transparency issues and data misuse have led to a growing mistrust of AI. That has increased calls for ethical audits of algorithms. Trust is mostly referenced in calls for trustworthy AI. Research organizations that are working with the ethics of AI are calling for ethical AI auditing, like EU High-Level Expert Group on Artificial Intelligence who made a draft, “Ethics Guidelines for Trustworthy AI”. (Brown et al. 2020). Today the problem is more “how algorithms audits are done” than “how audits could help building trust”. Trust is more like desired outcome from the auditing or driver for conducting AI auditing than a principle which can be audited.

Search engines are calling the need for algorithm audits because dependence and trust have undesirable effects on democracy. (Robertson et al. 2018.) For example, people use Google as their main fact checker tool for doing their own research as they are trusting the information which search engines provide. Floridi et al. (2018) believe that a good AI society needs to embed ethical principles in the default practices of AI, but they especially highlight explicability for ensuring public trust and understandability of the technology. Their goal is to develop AI technology in a way that secures people’s trust while serving public interest and strengthens social responsibility. That requires that society needs to develop a redress mechanism for harms inflicted, costs incurred, or other technology driven grievances. The mechanism needs to be accessible and reliable and involve clear and comprehensive allocation of accountability.

Many AI auditing drivers are closely related to trust. For example, provenance information needs to be reliable to be useful, but this also brings a few challenges. The accountability context is complex with given risks and incentives, inherent federation in terms of the mechanisms for capture and what is recorded and means for capture that are used impact reliability, validity, accuracy, usefulness and completeness. These all together raise issues of trust. (Singh et al. 2019.) The same is approached by Shneiderman (2020), who continues the matter by proposing 15 recommendations of three levels of governance aiming to increase reliability, safety and trustworthiness. According to them, these three are vital concepts to everyone involved in technology development. They bring benefits to individuals, organizations and society while they clarify who takes action and who is responsible. They suggest that external review organizations should use independent oversight methods as they can lead to the independent audits of products and services and trustworthy certification, create a trusted infrastructure to investigate failures, continuously improve systems, and gain public confidence. These independent oversight methods involve several different actors like government, auditing firms, professional organizations, society and insurance companies.

To get deeper on these models and mechanism, Harrison et al. (2020) investigated perceptions of fairness in ML models and compared their findings related to trust with Reed et al. (2016) study in which they investigated the relationship between trust and model properties. Reed et al. (2016) research lacked information about how differences in model properties across groups affect trust. Harrison et al. (2020) research showed that participant expressed more trust in human judges while Reed et al. (2016) participants favored algorithmic methods. Possible explanation for this is that Harrison et al. (2020) showed differences between model properties by racial group while Reed et al. (2016) focused only false positive rate, false negative rate and accuracy. Maybe participants would not have trusted algorithms if they were aware of difference across racial groups.’

### 3.2.8 Freedom & Autonomy

Last, freedom and autonomy was a recognized principle which featured in a total of six papers. This code included the terms freedom, autonomy, choice, self-determination, liberty and empowerment (Jobin et al. 2019). The idea of freedom and autonomy is that individuals should have a right to make their own decisions about treatment which they do or do not receive (Floridi et al. 2018). Reed et al. (2016) define autonomy as a fundamental right to have own choices rather than being forced to certain choices made for them. In everyday life autonomy could be impaired, for example, when someone lacks mental capacity to make decisions for themselves. With AI, people willingly give some of their decision-making power to the machines. The difficult part is to balance between the decision-making power we retain for ourselves and which we give to the AI systems. (Floridi et al. 2018.)

Floridi et al. (2018) report findings of AI4People, in which four documents dealt with principle of autonomy. First, The Montreal Declaration states that the control of autonomy systems and autonomy of all human beings should be the focus on development of AI. Group on Ethics in Science and New Technologies and UK House of Lords Artificial Intelligence Committee were along the same lines, arguing that people must be able to set their own standards and norms and AI systems should not impair that freedom. It is even said that the power to hurt, destroy or deceive humans should never be vested in ML systems. The Asilomar still states that it should be up to humans to choose how much they delegate their decision-making power to the AI systems. However, even though these documents express the topic similarly, they have slight difference ways approaching it balancing between beneficence and non-maleficence. Nonetheless, a central point is to the protect the value of human choice and their decisions and avoid the risk of delegating too much power to machines as humans should always retain to the power to decide which decisions to take.

Section 3.2.5 dealt with privacy. Data storage risks mentioned can further cause issues by potential consent violations during data collection. For instance, IBM collects their Diversity in Faces dataset from Flickr and while those images uploaded there are open and free to use, the individuals in the images might not have agreed to being included in a facial recognition dataset. (Raji et al. 2020.) Dulhanty et al. (2020) try to solve this issue with their experiment on a state-of-art system. Their objective is to define the impact of an individual's inclusion in face recognition training data on a derived system's ability to recognize them without taking a position for advocating technical improvements but more discussion about consent when it comes to face recognition. Results were quite concerning since no consent was sought or obtained in all datasets in their study which means that there do not exist any major open-source datasets with consent gathered from individuals.

Autonomy in ML systems is a problematic matter because the choices these systems make are based on what they have learnt. Principles or interpretations therefore do not guide actions and choices are often invisible to the users of technology. One very common example of ML autonomy problem is the choice which an ML system must make when self-driving vehicles are about to crash. Technology might have a situation where it must decide who will die and who will survive. ML technology is also subject to law and regulation which preserves autonomy and for these reasons' autonomy is one of the key drivers for auditing. (Reed et al. 2016.)

## 3.3 Actor-based approach

Third research question focuses on actor-based approach. One of the main concepts of auditing is to determine if auditing is made by external or internal actor. It is also interesting to survey to who are the stakeholders of AI auditing and who are proposed to conduct AI audits. Table 10 demonstrates in a cross-tabulation format how the sample was distributed between papers discussing internal and external auditing and between different actors. Top row shows to whom auditing papers were targeted and who conducted AI auditing. The left column shows if the paper dealt with internal or external auditing.

Table 10 Actor-based cross-tabulation

| | *Researchers* | *Systems development and deployment* | *Regulators* | *Individuals & companies* | *Auditors* | *Users* |
|---|---|---|---|---|---|---|
| ***Internal*** | | P1, P43 | P43 | | P26 | P11 |
| ***External*** | P4, P22, P25, P33, P39, P42, P45, P47, P48, P49 | P2, P4, P14, P30, P33, P38, P39, P41, P44, P45, P47, P48 | P10, P14, P34, P44, P45 | P6, P12, P14, P17, P25, P28, P38 | P2, P3, P10, P13, P14, P15, P23, P28, P37 | P20, P27, P39, P48 |
| ***Both*** | P31, P40 | | P35, P36 | P3, P35 | P18, P24 | P8, P36 |
| ***NA*** | P9, P21, P50 | P16 | | P29 | P7 | |

As we can see, external auditing is dominant here. This might be because ethical aspects of auditing interest more external than internal auditors. Sample was distributed between ethics-based and other category papers so that presumably affected why external actors emerged more. Another possible reason could be that it is easier to study ethics-based AI auditing from the external perspective or that internal auditing papers are published more in grey literature format. Between different actors the sample was distributed much more evenly. Most papers were focused for people who conduct AI auditing, people who are doing systems development and deployment or researchers.

Actor-based approach in AI auditing is divided into two parts: who are proposed to conduct AI audits and who are the stakeholders of AI auditing. In table 10, stakeholders of AI auditing are presented in top row. Researchers means that AI auditing was conducted by a researcher or that the paper was aimed for researcher purposes. If a clear subject to whom the audit was performed was not presented, the paper was defined as researcher category. Systems development and deployment, as its name suggest, indicates that AI auditing is conducted by a system developer or that AI auditing in paper is targeted for developers. Regulators are law enforcement or policy makers who either conduct AI audits or for which the auditing is performed. Individuals and companies refer to papers which were made for individuals and companies to use AI auditing either by conducting it or utilizing AI audits. Auditors specify papers which either create tools for auditors conducting AI auditing or papers where the role of auditor was not specified. Last, users refer papers in which users conduct AI auditing or AI auditing conducted in paper was made for users.

### 3.3.1 Who conducts AI auditing?

A fundamental matter about auditing in general is to determine whether auditing is made by internal or external actors. The same applies in AI auditing. In our sample, 5 out of 50 papers dealt with internal auditing, 33 dealt with external auditing, 8 dealt with both internal and external auditing and 6 papers could not be defined between internal or external auditing. A few possible reasons have already been presented earlier why external auditing is so dominant in this sample. It is also possibility that there are more external than internal actors conducting AI auditing or that external auditing is made or researched in more public way. Private companies, for example, might generate their own internal frameworks which are not public.

As stated earlier, audit is defined as "*an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures*" by IEEE Standard for software development. External algorithmic auditing has many similarities with external bug bounties, where hackers outside organization are paid to find bugs or vulnerabilities in the software. Audits increased public awareness of algorithmic accountability and other ethical standards as those revealed, for example, structural racisms and sexism in AI systems. Companies realized the need of understanding the social dynamics of their deployed systems' environments and internal auditing. (Raji et al. 2020.)

### Internal auditing

Mehrotra et al. (2019) argue that internal auditing methods are employed by service providers who use their own internal system information while external auditing rely only on publicly available information. For example, Facebook might audit its own algorithms internally, but external auditing methods are typically employed by third parties. In some cases, external vendors require components for internal auditing. Information security audit is such a required standard component of internal audit. Especially financial sectors and healthcare organizations combine internal and external auditing as they invite trusted third-party auditors to provide reports of the weaknesses, shortcomings, vulnerabilities etc. while also doing own internal auditing. (LaBrie et al. 2019.)

External auditors have no access to internal processes, but they can access model outputs. Intermediate models or training data are not included in external auditing as they are often protected as trade secrets. Internal auditing is therefore implemented to extend traditional external auditing. The goal with it is to evaluate how well the product or software fits the expected system behavior encoded in standards. Pre-deployment audits enables ethical intervention methods better than post-deployment audits. Identified gaps can be mapped with product teams. Internal auditors are employees of organizations, so they communicate findings primarily to internal audience. This enables organizations to make structural changes to auditability of processes and ethical standards and ultimately complementing external accountability, generating artifacts or transparent information that third parties can use for external auditing. (Raji et al. 2020.)

Internal decision-making algorithms are often hidden from the public. Internal auditors could, for example, use AI auditing for meeting legal standards or internal policies (Brown et al. 2020). However, companies often want to attach users in software development in a limited and controlled way. Users might be able to obtain responses from the algorithm for a few feature sets even though the internals of the model remain hidden. Therefore, individuals can audit deep models that make decisions on them. (Domingo-Ferrer et al. 2019.) Vendors and buyers might want to control ethical and reputational risks and other stakeholders might be interested in a general ethical assessment of an algorithm. This way the audit framework covers a wide range of different actors beginning from the internal regulatory or policies auditors to the users and other external auditors. (Brown et al. 2020.)

As presented in table 10, internal auditing is mostly conducted by systems development and deployment actors (Raji et al. 2020; Floridi et al. 2018), regulators (Floridi et al. 2018), selected auditors (Papakyriakopoulos et al. 2020) or users (Domingo-Ferrer et al. 2019; Cabrera et al. 2019). However, all of these come from the company's internal departments. Systems development and deployment auditors work with AI auditing mechanism to ensure that no unfair biases or other unwanted consequences happens. In addition, they develop solidarity mechanism to deal with severe risks in AI-intensive sectors. Structural vulnerability or bias disproportionately bring cost and harm to the systems and ethical requirements require that organizations give attention for auditing perspective and system developers are held into account for system compliance. (Raji et al. 2020; Floridi et al. 2018.)

Mehrotra et al. (2019) show a framework how organizations can internally audit online services conducted by internal auditors. They used three different internal auditing methods for measuring latent differences in user satisfaction using Bing as a case study. The results were partly similar than what Domingo-Ferrer et al. (2019) and Cabrera et al. (2019) come into in their research where users were part of conducting auditing. Domingo-Ferrer et al. (2019) presented a methodology that enables users to audit ML models that make decisions for them where the approach does not require that users have full access to the model, but they can still audit the system. Cabrera et al. (2019) introduce a system for users to audit the fairness in ML models. They built an interactive visual interface to help users explore the fairness.

### External auditing

Brown et al. (2020) presented an external audit instrument which could be used by regulators aiming to translate ethical analyses into practical steps. Regulators use ethical AI auditing to assess that algorithm fulfil legal standards or other internal guidelines. According to them, auditors doing audits for regulatory purposes should first identify stakeholders interested in regulatory agencies and then examine cases where the algorithm performs low on some metric that is highly relevant. Floridi et al. (2018) add to that self-regulatory codes of conduct for data and AI which helps to specify ethical duties and make sure that people understand the merits of ethical AI.

Auditing papers aimed for researchers were the most common in external auditing papers. Who actually conducts AI auditing was not clearly determined in those papers. Kyriakou et al. (2018), Barlas et al. (2020) and Barlas et al. (2019), for example, wanted to understand how image analysis algorithms work, how those treat people and how to develop ways to audit them. According to them, third party developers auditing is on the rise and algorithms are audited from the outside when full transparency is not possible. Two approaches they presented for external auditing were within-platform auditing and cross-platform auditing. Within-platform auditing means that the input it systematically manipulated to study how the resulting outputs differ. They give Sweeney (2013) as an example who tested racial bias in Google AdSense by manipulating names by their racial associations and compared the ads chosen by the algorithm. Cross-platform auditing is for detecting cases where a system is generally biased, for example, comparing different hotel booking platform ratings. (Kyriakou et al. 2018.)

Mittelstadt et al. (2016) suggest that data processors, external regulators or empirical researchers should conduct AI auditing for achieving explainability. Auditing is necessary to verify algorithms' correct functioning. Practical solutions require cooperation between researchers, developers and policymakers. Chen et al. (2018) show an audit study methodology where researchers study hiring practices and outcomes. In their research, they empirically do an algorithmic auditing on online job boards revealing algorithmic unfairness. Also, studies Barlas et al. (2020), Barlas et al. (2019), Singh & Hofenbitzer (2019) and Jiang & Vosoughi (2020) show how researchers can do algorithmic auditing with external API data. Singh & Hofenbitzer (2019) used Twitter data to audit cyberbullying and to create more accurate and fair cyberbullying detection algorithms. That shows how to audit an existing social network with selected algorithms. Accordingly, Jiang & Vosoughi (2020) audited the performance of toxicity of Twitter users using Perspective API. They show how with a free tool they can audit the data and consider the consequences of the behavior and gain insight into how a ML system performs across a range of inputs and parameters.

Systems development and deployment actors conducting AI auditing were also a highly represented group in external auditing papers. Clavell et al. (2020) provide insights how collaboration between developers and algorithmic auditors can lead to better technologies. Telefonica Innovacion Alpha developed an Algorithmic Audit REM!X to decrease the discrimination of protected groups by identifying and mitigating algorithmic biases. The problem is that 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 which identifies potential discriminations. There are both technical barriers and technical literacy that are causing troubles on understanding and adapting ML methods. With the help of researchers, developers can get tools to audit the outputs of cognitive services, understand the benefits and risks and making best choices of cognitive services. (Barlas et al. 2019.)

Auditing cognitive services is also the interest of Barlas et al. (2020). They write about developers, designers and researchers who are interested in incorporating auditing tools into their work and how stakeholders or users could use the developed tools for auditing. Formal auditing is done by developers as they have full access to the system, but third-party auditors could also do auditing, for example, through COMPAS system. Adding to this, LaBrie et al. (2019) are calling for ethical AI framework for AI development and deployment. Ethical audits are similar to security or accounting audits as trusted third-party auditors perform those. Purpose of auditing is to give the organization's goals and expectation of the AI algorithm.

When it comes to external auditing, the auditor is often either an independent auditor, an external auditor hired by a company, or an auditor hired by an outside source. However, professional auditors are for making use of the audits in practice and in policy. Raji et al. (2020) present datasets such as Face Recognition Vendor Tests, the Pilot Parliaments Benchmark and the IARPA Janus Benchmarks. They show how audits conducted on these datasets have heavily impacted FTP benchmarks and frameworks. Using their audit conducted with CelebSET, auditors could explore ethical concerns in current algorithmic auditing practices. Among same lines was Robertson et al. (2018) who conducted controlled algorithmic audit within Google Search to assess audience bias or Dulhanty et al. (2020) who performed audit to ArcFace, a state-of-the-art, open-source face recognition system. Robertson et al. (2018) described AI auditing methodology and what it takes to conduct AI auditing and how auditors can use tools provided by algorithmic auditing methodologies for assessing their output based on a controlled input. In addition, they mention that this is a great starting point for externally auditing the impacts, but broader frameworks would also need a user interface.

Multi-agent system auditing is made for answering the need of external and neutral auditing in an AI based recruitment processes. The aim is to reduce the discrimination in the job market. (Martinez & Fernandez 2019.) This is one example of a situation where external auditors can perform audits to improve systems. There is also a scenario where GDPR can enhance a proactive auditing by data controllers, since data controllers might sometimes ignore the fact that an algorithm, they are using has biases. Often the information related to ethical principles is unknown to designers or to data controllers who use the algorithms which is why GDPR requires to perform an audit of decision-making algorithms. This makes sure that data controllers who use the system and conduct the auditing know the technical and organizational measures and can correct factors which have caused errors. The key elements for the auditor are the creation of the algorithm, how the algorithm works and what data it needs and what are the expected outputs of the algorithm. With these elements taken into consideration the data controller can conduct an AI audit. (Malgieri & Comande 2017.)

One last group conducting AI auditing are the users of the system. AIF360 is an example of designed workflow which is made for users to go from a raw data to a fair model and thus create results. It gives education on the important issues in bias checking and mitigation and helps to select which algorithms to use and how to use them. (Bellamy et al. 2018.) Same is with the Algorithmic Equity Toolkit presented by Katell et al. (2020). It is intended for community members to organize and outreach, participating in public comment sessions and gatherings, and assessing the impact of technologies. It helps users to understand how ML works, determines if the system is driven by AI and it asks questions about the context which helps users to reach their assessments. In addition, Barlas et al. (2020) points out that AI tools must not only be open, but users must have the knowledge and skills to understand them. Possible users might be the developer or researcher who collect outputs or create new system, process or analysis in cognitive services.

Kulshrestha et al. (2017) characterized the fairness of the ranking algorithms in Twitter's data. Their auditing framework helps users to become more aware of biases in a search process. This way they can use the system in a more intelligent way and know that system outputs might be biased. In other words, auditing framework gives users more control over the bias and mechanism makes them more aware of bias issues. Mechanism for discovering these biases is also addressed by Cabrera et al. (2019) as they presented their

FAIRVIS visual analytics tool for users to apply domain knowledge and analyze performances of subgroups. Users do not need to have previous knowledge about the system for using the FAIRVIS to find bias issues of the system.

### 3.3.2 Stakeholders of AI auditing

This section focuses on the stakeholders of AI auditing. As well as on topic about who conducts an AI auditing, there are multiple different actors to whom AI auditing is for. Table 10 presents the main actors, divided between internal and external auditing. Main actors to whom papers target AI auditing tools or concepts for are researchers, systems developers and deployers, auditors, users, individuals and companies, and regulators. There are multiple different ways how to utilize AI auditing. It is important for different stakeholders to import AI auditing into their work which is why there are many actors who are connected to it. Indirectly AI auditing also affects more or less to us all, but for clarity, I focus directly on the targets which are presented in the papers.

**Researchers**

Third party developers need to understand how image analysis algorithms treat people and how to audit them. That is why studies Kyriakou et al. (2018) and Barlas et al. (2019), for example, do groundwork for developers about how to conduct AI auditing. Kyriakou et al. (2018) give insights of processes and awareness of possible harms and biases, connects the work with ongoing conversation and deals issues with fairness in algorithmic systems. These are all for helping external auditors to improve their auditing processes and for research community to show the difficulties of studying fairness in ML systems. Sandvig et al. (2014) complement this view by stating that audit studies are typically conducted by a researchers who are doing field experiments and those studies are often targeted for employers. However, Sandvig et al. (2014) aim wider as they outline design ideas for empirical researchers giving them guide and agenda to research algorithmic discrimination.

It is common that researchers do not specify to whom their study is targeted for. For example, Sulaimon et al. (2019) aim to adapt their method at the existing bias detections for ensuring fairness in ML systems. Their goal is to give access to auditing decision processes of ML systems and improve existing systems which are used by auditors and system developers. However, they aim at enabling autonomous software systems to make an unbiased decision rather than giving tools or methods for auditors or system developers. Nonetheless, their study is a groundwork for further investigation. Similarly, Hanna et al. (2019), Bellamy et al. (2018), Chen et al. (2018) and D'Amour et al. (2020) all create methods or methodologies for researchers or practitioners to generally improve AI auditing.

Singh & Hofenbitzer (2019) and Jiang & Vosoughi (2020) use Twitter data to audit social network features. They also aimed their audit study for researchers to demonstrate network characteristics and the need of the community to understand multimedia processing and its unique ethical considerations and to improve algorithmic fairness field. Researchers were the most common target to whom AI auditing studies were for in our sample. However, AI auditing is not made for researchers as they just try to improve the field. Barlas et al. (2020), for example, made publicly available methodology about understanding machine behaviors and AI auditing. The goal is to build a controlled auditing approach for everyone's use who might benefit from it but other than other that researchers are not specified. This reflects well the overall picture of papers targeted to researchers.

### System development and deployment

LaBrie et al. (2019) suggested a framework for algorithmic development and deployment where algorithmic applications go through an ethical algorithm audit. Purpose of the framework is to externally provide System D&D actors best procedures for conducting an ethical AI auditing. They state that authors and owners of the algorithms could contribute of the framework for enhancing and correcting the issues discovered during the audit. They also highlight trusted third-party auditors who perform AI audits to benefit their framework. Among the same lines were Barlas et al. (2019), who aimed to provide tools for developers and researchers to audit algorithms when the code audit is not available. They strive to help developers to understand the benefits and risks of CogS, help them audit the outputs of Cogs and hence help developers to make the best choices of Cogs benefiting their needs.

Raji et al. (2020) introduce a framework aimed for internal developers to help them audit algorithms that supports AI systems. According to them, there is an accountability gap which they aim to close in order to ensure audit integrity in system development and deployment. AI technology affects billions of users which is why there is an increasing interest in corporations and governments to develop AI auditing mechanism. Raji et al. (2020) focus on internal auditing because external audits are mostly conducted on models after deployments which means that system might have already been impacted users in negative way. With internal auditing framework, they present mechanisms which assist developers to meet ethical expectations and standards in AI systems. Developers operating with internal audits can therefore prevent potential negative consequences beforehand and abandon the development of AI technology which causes more risks than benefits. The SMACTR framework by Raji et al. (2020) is made for supporting this development of AI systems.

Responsibility is one of the main drivers in designing and developing AI systems. ML techniques, training data, ML outputs and the system context are all related to responsibility, which are driving developers to reach control and transparency expectations. It is important for people involved in developing ML systems to recognize and to avoid potential issues of responsibility. For this, many ML tools, services and standards are made widely available to assist and guide ML development and deployment processes. (Singh et al. 2016.) This is also addressed by Mittelstadt et al. (2016), who aimed to clarify the ethical importance of AI development and to identify areas of it. Explainability in AI auditing can be carried out by developers, and they aim to offer conceptual framework for ethical inquiries and development of algorithms.

One of the main aspects of AI auditing is to understand how algorithms treat people-related media. For this, Kyriakou et al. (2018) seek solutions to develop ways to audit them. They are especially targeted for third-party developers as third-party developing is on the rise since CogS, for example, is gaining popularity. However, there are many problems that developers might not expect when using APIs in development processes. Kyriakou et al. (2018) presented a gender inference, judgment tags and abstract inferences as behaviors which might interpret people in an unfair manner. Therefore, they present auditing tools for any developer to use and enable developers to be aware of different scenarios. However, Barlas et al. (2020) present an experiment where they examined the interdependence between algorithmic recognition of context and the depicted person’s gender. They made a publicly available code and reviewed auditing approaches for motivating the need to develop auditing procedures in opaque services like CogS. It is surveyed that developers do not have direct need for tools but rather aspirational motivation for them. It was also reported that ML tools must be open and interpretable to be transparent and that developers must have skills to understand them.

Audit tools are often made for protecting certain groups and finding ways for developers to collaborate between auditors in order to better technologies. Algorithm Audit of REM1X is an example of developed app which is made for developers to establish new procedures and safeguard in AI development so that they can answer the prompting needs for audit their algorithmic services and procedures and help it users to establish auditing practices. (Clavell et al. 2020.) Another toolkit aimed for developers is AI Fairness 360 by Bellamy et al. (2018). Purpose of this toolkit is to enable developers to make improvements for new algorithms and use it for performing benchmarking. There is ambiguity in fairness scientific and AI practitioner's community where AIF360 strives to give solutions and solve issues. AIF360 makes it easier for developers understand ethical metrics and foster further contributions and information sharing and hence deploy solutions in different industries. It offers education guidance and tutorials on important issues on bias detection and which algorithms to use.

In addition to tools, researchers propose recommendations and ethic codes for developing AI. Floridi et al. (2018) list ethical principles which developers should adopt in their AI development process in order to establish better AI society. They also offer set of recommendations to support this. Principles and recommendations adopted in AI development process serve all the stakeholders and increase public trust and acceptance of development process, for example, by bringing new abilities and skills in the scene and mitigating its impact on old procedures. According to them, thoughtfully developed AI system offers opportunities and improvements both human agencies, organizations and human life in general as developed audit mechanisms for AI systems identifies unwanted consequences and deals with several risks in AI sectors. For the same purpose, Grasso et al. 2020 presented its framework demonstrating how compounding accountability frameworks and domain-specific codes of ethics can help answering ethical expectations in systems which utilize AI. Grasso et al. (2020) framework is specially targeted for developers so that they can avoid unintended consequences when deploying ML systems. Amazon, for example, developed a resume sorting system which downgraded women. With the help of this framework, developers can recognize critical knobs of decision-making systems and apply a code of ethics into their work.

### Users

Some of the auditing tools and framework presented by researchers are aimed directly for users so that they can audit algorithms which affects them. Therefore, AI auditing is aimed for system users but also for the whole user community. As discussed earlier, open-source algorithms would be a simple solution for algorithms to be transparent but for industrial or intellectual property reasons this cannot always be the case. This is why Domingo-Ferrer et al. (2019) present approach where users to whom the AI system affects can in a collaborative way make a rule-based approximation of the model underlying the decision algorithm. This allows users to go against opaque decisions-making systems. Even though the approach encourages users to serve their own benefit, it will also benefit the entire user's community. This is called "co-utile situation".

FAIRVIS is one example of a tool which is targeted directly to the users. With the help of FAIRVIS, users can audit ML models by themselves. Therefore, researchers aimed AI auditing for the users but made it so that the users can explore subgroup performances, apply domain knowledge to develop and investigate known subgroups and explore similar subgroups. This way they can visualize fairness and performance metrics and compare how their performance differs to others. (Cabrera et al. 2019.) Similarly, Algorithm Equity Toolkit by Katell et al. (2020) is made for users so that they can understand and recognize AI systems and the potential harms better and hence hold policymakers accountable. Community members, grassroots organizations, and members of the general public are the main users to whom Algorithm Equity Toolkit is targeted. Main ways how it is supposed to help users are: 1) Determine whether a system relies on AI capabilities, 2) asking questions for advocating the social context of a system so that communities can demand answers from policymakers and 3) increase understanding about how ML systems work and how they fail.

Kulshrestha et al. (2017) studied search engines in order to increase transparency and decrease discrimination for users. People rely heavily on search engines, for example, when they are gathering new information. Therefore, Kulshrestha et al. (2017) propose a framework which enable users to be more aware of the potential biases in search results and gives users more control. For the same transparency and bias problems Brown et al. (2020) developed their audit instrument. Purpose of it is to increase public trust of AI systems and reduce misuse of data. It serves the interest of all the users and stakeholders affected by algorithms. For example, users might not know what data is collected about them and how long it is stored. However, even thought the motivation of the framework was to serve the interest of people affected by algorithm, the proposed instrument is also aimed for regulators to use. Many questions about how audits should look like when they are intended to be used by regulators are in the air and Brown et al. (2020) aim to suggest an auditing instrument which transforms those ethical analyses into practical steps for regulators. The focus is on negative impacts, especially on how an algorithm meets legal standards or internal policies, since regulators are mostly interested in those.

**Regulators**

As AI is so powerful, it needs to be regulated. Good regulation also allows AI to reach its potential as it reduces fear, ignorance and misplaced concerns about it. Widely accessible mechanisms of regulation increase public acceptance and adoption of AI technologies. For this, Floridi et al. (2018) studied if the current regulations are corresponding with the ethical principles and the current state of AI systems. They encourage to include ethical, legal and social considerations in AI projects so that AI projects answer both ethics and policy calls. They also mention self-regulatory codes of conduct as many current techniques can be constrained through it. Likewise, Mittelstadt (2019) noted that AI companies and developers are committing themselves to ethical principles and self-regulation codes which might affect that policy-makers do not pursue new regulations. However, organizations are required to work within strict regulatory frameworks. Processing of personal data, for example, follows strict criteria. According to Mittelstadt (2019), a unified regulatory framework does not exist in an AI field. Therefore, they aim to provide legal mechanism for regulators and seek to provide direction for regulation framework for AI development.

The EU GDPR prompts companies and organizations to audit their services. This also gives organizations stronger safeguards and setting for keeping bar high for decision-making systems and protocols for addressing algorithmic fairness, accountability, and transparency. (Clavell et al. 2020.) Concerns of algorithmic discrimination towards individuals and groups were one of the biggest motivations for the GDPR. It is specially focused on protecting personal data, but it also gives legislation to address the effects of algorithmic decision making. These pave the way for third party inspections of AI auditing. Regulators can discover and reduce discrimination and improve accountability via auditing. However, GDPR does not specify who should perform audits. Private auditors or public monitors could both benefit from GDPR in their own ways. Costs of auditing and roles of companies for helping the auditing process is also a topic of discussion. For these questions legal and formal definitions are needed for regulators. (Goodman 2016.) Regulators need to have better mechanisms and methodologies to address auditing issues. Clear policies and standards, which GDPR aims to bring, is a significant area for further work.

### Auditors

Auditors are naturally a key target for AI auditing studies. Whether it is by creating a framework or system architecture for better auditing practices (LaBrie et al. 2019; Kim et al. 2019; Martinez & Fernandez 2019), studying auditing elements of social networks (Robertson et al. 2018; Toapanta et al. 2020) or ethical problems related to facial recognition systems, the auditor is in the center. That is why many researchers aim their auditing studies for the auditors. Auditors could benefit an ethical framework for AI audits and answer better organizations goals and expectations (LaBrie et al. 2019). This could also be adopted in proposed multi-agent system architecture, by Martinez & Fernandez (2019), which aims to reduce discrimination in job markets. They aim to automatize parts of AI auditing in HR. This will of course benefit people in the job market, but it will also affect highly in auditors' tasks. The multiaccuracy framework by Kim et al. (2019) also helps auditors to identify specific subgroups if the system if biased. If the predictor model makes mistakes, multiaccuracy framework will help auditors to identify those mistakes and why those mistakes happen and produce examples of inputs where the predictor is erring significantly.

As discussed earlier, search engines and other social networks highly affect people in general. This was the motivation in Robertson et al. (2018) study, in which they aimed to conduct a controlled algorithm audit of partisan audience discrimination and personalization within Google Search. This type of audit methodologies provides useful tools for AI auditors and benefit both individual information seekers and the whole society. Such audits will grow value and should be conducted regularly. Similar problems were addressed by Toapanta et al. (2020) to determine cyberbullying in social networks. They argue that social networks lack auditing methods which they aim to solve by providing protype for auditors to perform audits on social networks. Their prototype will elaborate auditing operations which helps auditors to carry out AI auditing accurately in shorter time. Also, Sapiezynski et al. (2019) made a novel metric, the Viable-Λ test, for auditors to answer questions about whether there exists a distribution of user attention such that output of search engine is group fair and what is the parameterization of the distribution. This can be used by internal auditors so that they can ensure that they deliver fair results.

Another recurring theme of AI auditing discussed earlier is face recognition. Ethical issues in facial processing technology cause many harms to people if they are dealt unfairly. For this, Raji et al. (2020) aimed to highlight different ethical tensions which auditors need to be aware of. Their work will modify current auditing processes, develop new norms and help auditors to clarify limit of the audit scope as auditors needs to live up with the ethical ideals. Barlas et al. (2019) and Dulhanty et al. (2020) also studied harms cause by face recognition systems to the individuals and how to reduce it. Barlas et al. (2019) approach it by pointing out that tags produced by algorithms might not be fair and users should not be responsible for tagging their own images as technology might not be there yet. This puts also auditors in a position where they should focus on outputs of image tagging services. Likewise, Dulhanty et al. (2020) audited open-source face recognition system, ArcFace, for clarifying the impact of an individual's inclusion in training data on a derived system's ability to recognize them. People are uncomfortable that their faces are being used in technologies for the surveillance purposes. Dulhanty et al. (2020) analysis provides an alternative option for task-based auditing which shapes the algorithmic auditing of commercial face analysis applications.

**Individuals and companies**

As has been pointed out in many passages, motivation for ethical AI auditing is mostly on individuals, groups or companies. Whether the auditing tools, practices or frameworks where targeted for, for example, developers or auditors, the main object was to make sure fair, unbiased and transparent treatment towards people in general. Auditing algorithms is beneficial for both individuals and businesses. Legibility concept, which means combining transparency and comprehensibility, provides individuals convenience to understand functionality, impacts and the consequences of decision-making systems which is also provided in the GDPR provisions. (Malgieri & Comande 2017.) GDPR also offers a framework for automated individual decision-making (Clavell et al. 2020). Lack of readability and legibility have been the problem of algorithms when directly concerning individuals. Individuals deserve to be informed about the existence and logic about the system functionality and the specific decisions of decision-making algorithms. This can be reached by legibility-by-design systems led by GDPR. (Malgieri & Comande 2017.) However, it must be mentioned that GDPR does not define algorithmic discrimination or differentiate between disparate impact and disparate treatment. Disparate impact, which is GDPR's focus, means that neutral practices disadvantage special categories. Disparate treatment, in turn, means that an individual or a group receives unfavorable treatment based on any special categories which can be eliminated via mechanism introduced by the GDPR. (Goodman 2016.)

Human-centered AI is a widely discussed topic. AI auditing aims to limit risks and increase the benefits of it towards individuals and organizations. As discussed in section 3.2.2, the intention is to design human-centered AI which are reliable, safe and trustworthy, which in turn brings benefits to individuals. It will enable organizations to translate ethical principles into practices by modifying organizations structure in different levels. It allows, for example, a safety culture, trustworthy certification by external reviews and reliable systems for development teams. (Shneiderman 2020.)

Issues with bias has been one of the most discussed themes in every section. Papakyriakopoulos et al. (2020) showed how biases in word embeddings result algorithmic discrimination towards social groups and individuals. Therefore, frameworks for bias detection in concrete algorithmic applications of the embeddings need to be developed, quantify their impacts on individuals and mitigate the bias so that individuals do not get negatively influenced or discriminated against. This is possible through AI auditing as it helps to understand issues and identify needed measures for fair algorithms. (Papakyriakopoulos et al. 2020.) This kind of bias issues were also discussed earlier related to Google search problems. Billions of people are affected from decisions made by algorithms in online platforms. It influences heavily on individuals but also the whole society, culture and politics. That is why AI auditing is necessary for so many people. (Robertson et al. 2018.)

There are both direct and indirect discrimination towards individuals. Indirect discrimination means that even if the sensitive user features are not used by the system, it still has a correlation to the output of the system. This might cause disparate on the individuals being ranked. Direct discrimination means that user features are explicit used by the system when ranking people. These affect people in general, for example, in mentioned search engines, but also individuals in different occasions. Hiring discrimination was the focus of Chen et al. (2018) as they studied how discrimination have an influence on the candidates that are selected to fill open positions. They noted that audit studies are key tools for studying hiring discrimination. Buolamwini & Gebru (2018) share these thoughts. They argue that AI system which is not even trained to perform tasks like who is hired, fired or granted a loan, can be used in a pipeline to perform actions considering individuals. This may cause that someone is wrongfully accused on something or gets treated unfairly. For all these issues concerned, AI auditing provides increased demographic and phenotypic transparency and accountability in artificial intelligence.

# 4 Discussion

The present study performed a systematic literature review of ethics-based AI auditing to better understand the field and to identify special aspects that merit further discussion. For this purpose, four research questions were formulated. **RQ1** was aimed to identify primary ethical principles in ethics-based AI auditing literature. Detailed summarization of ethical principles can be seen in tables 9 and 10. **RQ2** was aimed to identify key drivers and dimensions of ethics-based AI auditing. This was carried out by connecting identified ethical principles with AI auditing. This way I intended to solve what kind of challenges or issues are driving the field, why these ethical drivers are important and how those are affecting the AI auditing. **RQ3** addressed the stakeholders of AI auditing. The main targets were identified and the importance and the impact of AI auditing for them was analyzed. **RQ4** explored actors who are proposed of conduct AI audits. As with RQ3, main actors for conducting AI auditing were listed and their roles and responsibilities were analyzed.

## 4.1 Theoretical contributions

This study has four key contributions to the literature on ethics-based AI auditing. The first contribution of this study was a detailed review of the current literature on ethics-based AI auditing. This included descriptive data of top publishers, main tools and frameworks, publication years and research methods. This structured approach helped to analyze the current state of the ethics-based AI auditing and showed the direction where the field is going. The review showed that the trend of publications is increasing every year and most of the publications are published in conference proceedings. Approaches for the ethics-based AI auditing challenges are evenly distributed as different research methods were used almost the same. Also, the division between new frameworks for guiding AI auditing and new tools for conducting AI auditing was relatively even.

The second contribution of this study was that I showed the main principles which appeared in the ethics-based AI auditing literature. Findings displayed that by far the majority of the papers considered fairness related issues, followed by transparency, responsibility and non-maleficence. However, definitions of the concepts vary a lot. For example, transparency can be seen as a wider concept of accountability being a concept of accountability or transparency can be seen as a separate concept for the accountability.

This study contributed to harmonizing definitions of different principles. Most of the principles have no unique definition or they are not defined at all in the papers. The review brought together different definitions and highlighted the concepts they used. This could provide a contribution for future researchers if they aim to conceptualize AI auditing.

The search displayed emerging convergence around certain ethical principles. Fairness and bias, for example, were used in a headline in total of 24 papers which might suggest that for ethical auditing of AI the biggest concerns are around these concepts. One explanation could be that the end goal for ethical AI auditing is often to ensure that AI treats people right, and most of the AI ethical problems are that AI treats people unfairly. Review also showed that papers addressing non-maleficence are significantly higher than papers addressing beneficence which could imply that ethical auditing of AI is more concerned about preventing harms than highlighting benefits. For further research it is also important to note that sustainability, dignity and solidarity concepts were not addressed in the papers. Concerns related to individuals were brought to the center which might be a reason why, for example, environmental aspects were not taken into account. The second reason could be that dignity or solidarity were not considered as relevant concepts for approaching humanitarian challenges as, for example, fairness or bias.

The third contribution was to link ethical principles with AI auditing to identify drivers and dimensions of the field. This helped to understand real-world impacts of ethics-based AI auditing. I identified whether something is audited directly or whether achieving ethics is a desired outcome. The division is not unambiguous as some of the principles can be seen as a driver or a dimension. However, most of the times ethics-based AI auditing aims to identify and prevent harms. This means that ethical AI auditing principles are not necessarily audited directly, but they can be reached via AI auditing, therefore, beneficence and non-maleficence can be seen as drivers of AI auditing. Also, fairness relates issues are often drivers even though they can be audited directly. Trust and freedom are also clearly more of a desired outcomes than a directly audited principles.

Some tools and framework are directly developed for the purpose to audit, for example, transparency or accountability issues. Many new AI technologies have serious ethical concerns. Image analysis technology, for example, is growing in popularity, but from ethical point of view it has challenges with negative biasing and transparency. Auditing is necessary to verify correct functions in different decision-making algorithms. Hence, transparency, responsibility, fairness and privacy can be seen as dimensions of ethics-based AI auditing.

The fourth contribution of this study is the outlined actors and targets of AI auditing. Key actors conducting AI auditing were identified and their roles were systematically mapped. The division between internal and external AI auditing was interesting as external AI auditing was so dominant. However, this could be due to ethics-based paper division. Nonetheless, there is a fundamental difference whether an organization audits their processes themselves or whether it is done by an external actor. Study showed that more academic research is needed for the internal auditing purposes. It was also interesting aspect to review the stakeholders of AI. In some cases, auditing was conducted for organizations purposes, but most of the cases AI auditing aimed to benefit individuals and groups outside organizations. A division also took places between whether AI auditing was conducted for auditors' own purpose or whether it was conducted for benefit others.

The field of AI auditing is moving rapidly. It is difficult to anticipate in which direction the development is going. It might consolidate in technical solutions or ethical AI auditing might end up being a guideline for AI auditing. It is also interesting to see in which direction regulation is going and how companies must conduct internal AI auditing and how much they have to participate in external AI auditing. This study was important starter as AI auditing is still non-established sector which needed more mapping of fundamental questions.

## 4.2 Limitations

Limitations of this study are that it only focuses on the ethics-based side of AI auditing, and the division of ethical papers are based on existing guidelines. Therefore, the papers were screened and analyzed on the basis of the principles given, but it did not aim to recognize new or unidentified areas. Future studies could add other sectors to the analysis or aim to identify principles which are not yet considered. Also, the inclusion of literature published considered only academic publications. Sources from grey literature could be beneficial to add in future studies.

This thesis also did not consider auditing process profoundly nor study internal auditing practical implementation or position in the organizations. The viewing point was ethics-based which may divert attention from other perspectives. For example, risk and controls matrix by KPMG (2018) go through supplier management, business process and other business-oriented perspectives which are different from the ethics-based AI auditing perspectives.

## 4.3 Future work

For future research directions it would be interesting to review a continuum of AI auditing literature from other aspects than ethics. This thesis did not address technical or legal point of view which would be interesting direction to review further. Also, social and technical interface is interesting aspect to research. What is the role of people and what is the role of software and how those affects each other would be interesting to investigate further. It would be also beneficial to research how much there are auditing papers which use AI in auditing to investigate how wide is the whole area.

# 5 Conclusion

In this study I aimed to understand the current state of ethics-based AI auditing. To address this, an SLR was performed on four databases: Scopus, Web of Science Core Collection, IEEE Xplore and AMC Digital Library. From the sample of 50 ethics-based AI auditing articles, I was able to synthetize the most important ethical principles in ethics-based AI auditing literature and linking them with drivers and dimensions of AI auditing field. In addition, I recognized the most important stakeholders of AI auditing and actors who are proposed to conduct AI audits. The review and the search process followed the PRISMA guidelines, which ensures that review is thoroughly conducted.

The findings were used to summarize the existing ethics-based AI auditing literature. The review highlighted several key characteristics of the ethics-based AI auditing. Fairness, transparency, non-maleficence and responsibility were the most common ethical principles, following with privacy, beneficence, freedom and autonomy and trust. The review concludes that fairness, non-maleficence, beneficence and trust are key drivers of ethics-based AI auditing while fairness, transparency, responsibility and privacy are key dimensions of it.

The findings also addressed the most important stakeholders of AI auditing. Those were researchers, system developers, regulators, auditors, users and individuals and companies. Their roles and responsibilities varied depending on their position meaning whether they were proposed to conduct AI audits or whether they were meant to benefit of it. Stakeholders who were proposed to conduct AI audits were identified to be mostly external actors. Addressing these elements in future studies can further develop the understanding of the AI auditing field and, for example, further clarify the differences between the technical system and the roles of the people and their interaction.

## References

AI HLEG (2019) *Ethics guidelines for trustworthy AI.* Independent high-level expert group on artificial intelligence set by European Comission. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.>, retrieved 21.2.2021.

Barlas, P. – Kleanthous, S. – Kyriakou, K. – Otterbacher, J. (2019) What Makes an Image Tagger Fair?. *Conference on User Modeling, Adaptation and Personalization*, 95–103.

Barlas, P. – Kyriakou, K. – Guest, O. – Kleanthous, S. – Otterbacher, J. (2020) To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proceedings of the ACM on Human-Computer Interaction*, Vol 4 (3), 1–31.

Barlas, P. – Kyriakou, K. – Kleanthous, S. – Otterbacher, J. (2019) Social B(eye)as: Human and Machine Descriptions of People Images. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*, Vol. 13 (1), 583-591.

Bellamy, R. – Dey, K. – Hind, M. – Hoffman, S. – Houde, S. – Kannan, K. – Lohia, P. – Martino, J. – Mehta, S. – Mojsiloviv, 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. *Computer Science*.

Bird, S. – Dudik, M. – Edgar, R. – Horn, B. – Lutz, R. – Milan, V. – Sameki, M. – Wallach, H. – Walker, K. (2020) Fairlearn: A toolkit for assessing and improving fairness in AI. *White paper published by Microsoft*.

Black, E. – Yeom, S. –Fredrikson, M. (2020) FlipTest: Fairness Testing via Optimal Transport. *In Conference on Fairness, Accountability, and Transparency*, 111–121.

Brown, S. – Davidovic, J. – Hasan, A. (2020) The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, Vol. 8 (1).

Buolamwini, J. – Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Butcher, R. – Beridze, C. (2019) What is the State of Artificial Intelligence Governance Globally?. *The RUSI Journal,* Vol 164 (5-6), 88–96.

Cabrera, A. – Epperson, W. – Hohman, F. – Kahng, M. – Morgenstern, J. – Chau, D.H. (2019) FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *IEEE Conference on Visual Analytics Science and Technology.*

Carrier, R. – Brown, S. (2021) Taxonomy: AI Audit, Assurance & Assessment. <https://static1.squarespace.com/static/5ff3865d3fe4fe33db92ffdc/t/60329e0a4cfbaa172691f7e6/1613929999802/Taxonomy+of+AI+Audit+%282%29.pdf>, retrieved *25.2.2021.*

Chen, L. – Hannak, A. –Ma, R. – Wilson, C. (2018) Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.

Clarke, L. (2021) AI auditing is the next big thing. But will it ensure ethical algorithms?. <https://techmonitor.ai/technology/ai-and-automation/ai-auditing-next-big-thing-will-it-ensure-ethical-algorithms>, retrieved *15.7.2021.*

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

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

D’Amour, A. – Baljekar, P. –Srinivasan, H. – Sculley, D. – Atwood, J. – Halpern, Y. (2020) Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. *In Conference on Fairness, Accountability, and Transparency*, 525–534.

Davies, D. – Jindal-Snape, B. – Collier, C. – Digbya, R. – Haya, P. – Howe, A. (2013) Creative learning environments in education—A systematic literature review. *Thinking Skills and Creative*, Vol 8, 80–91.

Deloitte (2020) IT Auditing: The process involved and its importance in today’s business. <https://www2.deloitte.com/mt/en/pages/risk/articles/mt-risk-article-it-auditing-process.html>, retrieved *15.7.2021.*

Dignum, V. (2019) *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way.* Springer, Switzerland.

Domingo-Ferrer, J. – Perez-Sola, C. – Blanco-Justicia, A. (2019) Collaborative Explanation of Deep Models with Limited Interaction for Trade Secret and Privacy Preservation. *Companion Proceedings of The 2019 World Wide Web Conference,* 501–507.

Dulhanty, C. – Wong, A. (2020) Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification. *In 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 244–250.

Epstein, Z. – Payne, B. – Shen, J. – Hong, B. – Felbo, A. – Dubey, M. – Groh, M. – Obradovich, N. – Cebrian, M. – Rahwan, I. (2019) TuringBox: An Experimental Platform for the Evaluation of AI Systems. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence,* 6826–5828.

Fernandez, C. – Fernandez, A. (2019) AI in Recruiting. Multi-Agent Systems Architecture for Ethical and Legal Auditing. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.*

Floridi, L. – Cowls, J. – Beltrametti, M. – Chatila, R. – Chazerand, P. – Dignum, V. – Luetge, C. – Madelin, R. – Pagallo, U. – Rossi, F. – Schafer, B. – Valcke, P. – Vayena, E. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines*, 689–707.

Gaumond, E. (2021) Artificial Intelligence Act: What Is the European Approach for AI?. < https://www.lawfareblog.com/artificial-intelligence-act-what-european-approach-ai>, retrieved *9.8.2021.*

Goodman, B. (2016) A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection. *29th Conference on Neural Information Processing Systems*.

Google (2020) What-If-Tool. Partnership on AI. <https://pair-code.github.io/what-if-tool/index.html>, retrieved 6.9.2021.

Grasso, I. – Russell, D. – Matthews, A. – Matthews, J. – Record, N. (2020) Applying Algorithmic Accountability Frameworks with Domain-specific Codes of Ethics: A Case Study in Ecosystem Forecasting for Shellfish Toxicity in the Gulf of Maine. *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, 83–91.

GreyNet (2013) GreyNet: Grey Literature Network Service.<http://www.greynet.org/>, retrieved 5.6.2021.

Hanna, A. – Denton, E. – Smart, A. – Smith-Loud, J. (2019) Towards a Critical Race Methodology in Algorithmic Fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512.

Harrison, G. – Hanson, J. – Jacinto, C. – Ramirez, J. – Ur, B. (2020) An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. *In Conference on Fairness, Accountability, and Transparency*, 392–402.

Hinson, G. (2007) The State of IT Auditing in 2007. *The EDP Audit, Control, and Security Newsletter*, Vol. 36 (1), 13–31.

Hu, X. – Rousseau, R. – Chen, J. (2011) On the definition of forward and backward generations. *Journal of Informetrics*, Vol. 5 (1), 27–36.

Iden, J. – Eikebrokk, T. (2013) Implementing IT Service Management: A systematic literature review. *International Journal of Information Management*, Vol 32 (3), 512–523.

Ilvento, C. – Jagadeesan, M. – Chawla, S. (2020) Multi-Category Fairness in Sponsored Search Auctions. *In Conference on Fairness, Accountability, and Transparency*, 348–358.

Jiang, J. – Vosoughi, S. (2020) Not Judging a User by Their Cover: Understanding Harm in Multi-Modal Processing within Social Media Research. *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, 6–12.

Katell, M. – Herman, B. –Binz, C. – Young, M. – Guetler, V. – Raz, D. – Dailey, D. – Tam, A. – Krafft, P.M. (2020) Toward Situated Interventions for Algorithmic Equity: Lessons from the Field. *In Conference on Fairness, Accountability, and Transparency*, 45-55.

Katell, M. – Herman, B. –Binz, C. – Young, M. – Guetler, V. – Raz, D. – Dailey, D. – Tam, A. – Krafft, P.M. (2020) Toward Situated Interventions for Algorithmic Equity: Lessons from the Field. *In Conference on Fairness, Accountability, and Transparency*, 45-55.

Kearns, M. – Neel, S. – Roth, A. – Wu, Z.S. (2019) An Empirical Study of Rich Subgroup Fairness for Machine Learning. *In FAT* ’19: Conference on Fairness, Accountability, and Transparency.*

Kim, M. – Ghorbani, A. – Zou, J. (2019) Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *In AAAI/ACM Conference on AI, Ethics, and Society,* 247–254.

KPMG (2018) A Risk and Controls Matrix. < https://pair-code.github.io/what-if-tool/index.html >, retrieved 6.9.2021.

Kroll, J. – Huey, J. – Barocas, S. – Felten, E. – Reidenberg, J. – Robinson, D. – Yu, H. (2016) Accountable algorithms. *University of Pennsylvania Law Review*, Vol. 165 (3), 633–705.

Kulshrestha, J. – Eslami, M. –Messias, J. – Zafar, M. – Ghosh, S. – Gummadi, K. – Karahalios, K. (2017) Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432.

Kyriakou, K. – Barlas, P. – Kleanthous, S. – Otterbacher, J. (2019) Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*, Vol. 13 (1), 313–322.

LaBrie, R. – Steinke, G. (2019) Towards a Framework for Ethical Audits of AI Algorithms. *Twenty-fifth Americas Conference on Information System*, 33-44.

Leyer, A. – Schneider, S. (2021) Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers?. *Business Horizon,* Vol. 64 (5), 711–724.

Mackenzie, C. (1998) Ethical Auditing and Ethical Knowledge. *Journal of Business Ethics*, Vol. 17 (13), 1395–1402.

Magee, K. (2021) IT auditing and controls – planning the IT audit. < https://resources.infosecinstitute.com/topic/itac-planning/>, retrieved 7.7.2021.

Mahood, Q. – Eerd, D. – Irvin, E. (2013) Searching for grey literature for systematic reviews: challenges and benefits. *Research Synthesis Methods,* Vol. 5 (3), 221–234.

Malgieri, G. – Comande, G. (2017) Why a Right to Legibility of Automated Decision-Making Exist in the General Data Protection Regulation. *International Data Privacy Law*, Vol. 7 (4), 243–265.

Margues, R. – Santos, C. (2017) Research on continuous auditing: A bibliometric analysis. *12th Iberian Conference on Information Systems and Technologies.*

Marquez, A. – Schneider, S. (2021) Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers?. *Business Horizon,* Vol. 64 (5), 711–724.

Mehrabi, N. – Morstatter, F. – Saxena, N. – Lerman, K. – Galstyan, A. (2019) A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys,* Vol. 54 (6), 1–35.

Mehrotra, R. – Anderson, A. –Diaz, F. – Sharma, A. – Wallach, H. – Yilmaz, E. (2017) Auditing Search Engines for Differential Satisfaction Across Demographics. *Proceedings of the 26th International Conference on World Wide Web Companion*, 626–633.

Microsoft (2020) Fairlearn: A toolkit for assessing and improving fairness in AI. < https://resources.infosecinstitute.com/topic/itac-planning/>, retrieved 7.7.2021.

Mittelstadt, B. – Allo, P. – Taddeo, M. – Wachter, S. – Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society*, Vol 3 (2), 1–21.

Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, Vol 1 (11), 501–507.

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

Obermeyer, Z. – Powers, B. – Vogeli, C. – Mullainathan, S. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, Vol 366 (6464), 447–453.

Omoteso, K. (2012) The application of artificial intelligence in auditing: Looking back to future. *Expert Systems with Applications*, Vol. 39 (9), 8490-8495.

Ongsulee, P. (2017) Artificial Intelligence, Machine Learning and Deep Learning. *15th International Conference on ICT and Knowledge Engineering,* 1–6.

Papakyriakopoulos, O. – Serrano, J. – Hegelich, S. – Marco, F. (2020) Bias in Word Embeddings. *In Conference on Fairness, Accountability, and Transparency*, 446–457.

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

Raji, I. – Mitchell, M. – Buolamwini, J. – Lee, J. – Gebru, T. – Denton, E. (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for

Internal Algorithmic Auditing. *In Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Raji, I. – Mitchell, M. – Smith-Loud, J. – Smart, A. – Gebru, T. – Theron, D. – Hutchinson, B. – Barner, P. (2020) Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44.

Reed, C. – Kennedy, E. – Silva, S. (2016) Responsibility, Autonomy and Accountability: legal liability for machine learning. *SSRN Electronic Journal.*

Robertson, R. – Jiang, S. –Joseph, K. – Friedland, L. – Lazer, D. – Wilson, C. (2018) Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction*, Vol 2, 1–22.

Rosthorn, J. (2000) *Business Ethics Auditing — More Than a Stakeholder's Toy.* Springer, Dordrecht.

Sandvig, C. – Hamilton, K. – Karahalios, K. –Langbort, C. (2014) Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *A preconference at the 64th Annual Meeting of the International Communication Association*.

Sapiezynski, P. – Zeng, W. – Robertson, R. – Mislove, A. – Wilson, C. (2019) Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. *Companion Proceedings of The 2019 World Wide Web Conference*, 553–562.

Scheuerman, M. – Wade, K. – Lustig, C. – Brubaker, J. (2020) How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, Vol 4 (1), 1–35.

Selcuk, A. (2019) A Guide for Systematic Reviews: PRISMA. *Turkish Archives of Otorhinolaryngology*, Vol 57 (1), 57-58.

Shneiderman, B. (2020) Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems,* Vol 10 (4), 1–31.

Singh, J. – Cobbe, J. – Norval, C. (2019) Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access,* Vol. 7, 6562–6574.

Singh, J. – Walden, I. – Crowcroft, J. – Bacon, J. (2016) Responsobility & Machine Learning: Part of a process. *SSRN Electronic Journal*.

Singh, V. – Hofenbitzer, C. (2019) Fairness across Network Positions in Cyberbullying Detection Algorithms. *Fairness across Network Positions in Cyberbullying Detection Algorithms*, 557-559.

Song, C. – Shmatikov, V. (2019) Auditing Data Provenance in Text-Generation Models. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 196–206.

Sulaimon, I. – Ghoneim, A. – Alrashoud, M. (2019) A New Reinforcement Learning-Based Framework for Unbiased Autonomous Software Systems. *8th International Conference on Modeling Simulation and Applied Optimization.*

Tan, S. – Caruana, R. –Hooker, G. – Lou, Y. (2018) Detecting Bias in Black-Box Models Using Transparent Model Distillation. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Tandon, A. – Dhir, A. – Almugren, I. – AlNemer, G. – Mäntymäki, M. (2021) Fear of missing out (FoMO) among social media users: a systematic literature review, synthesis and framework for future research. *Internet Research*, Vol 31 (3), 782–821.

Toapanta, S. – Monar, J. – Gallegos, L. (2020) Prototype to Perform Audit in Social Networks to Determine Cyberbullying. *World Conference on Smart Trends in Systems, Security and Sustainability*, 145–153.

Virovere, A. – Rihma, M. (2008) Ethics Auditing and Conflict Analysis as Management Tools. *Working paper*, 67–79.

Zhang, H. – Gao, L. (2019) Shaping the Governance Framework towards the Artificial Intelligence from the Responsible Research and Innovation. *IEEE International Conference on Advanced Robotics and Its Social Impacts*.